Master Thesis

Identification of Gene Transfer Events in Viruses

Saskia Metzler

First Reviewer: Dr. Olga Kalinina
Second Reviewer: Prof. Dr. Dr. Thomas Lengauer

Monday 9th December, 2013
Statutory Declaration

I declare that I have developed and written the enclosed thesis entirely by myself and have not used sources or means without declaration in the text. Any thoughts or quotations which were inferred from these sources are clearly marked as such. This thesis was not submitted in the same or in a substantially similar version, not even partially, to any other authority to achieve an academic grading.

Saarbrücken, Monday 9th December, 2013

Saskia Metzler
Viruses are particles that infect other organisms in order to replicate. They are highly diverse and can infect all kinds of cellular organisms. As a result of their close interaction with the host and their ability to switch between hosts, they are assumed to be one of the driving forces of evolution: They enable the exchange of genetic material between organisms. Both, donation to the host and take-up of genetic material from the host have been observed in different viral species. In this work, two systematic analyses of gene transfer in all to date known viruses are presented.

The first analysis is targeted at the identification of common proteins in unrelated viruses. Conducting a BLAST search, we find proteins in viral species with similarity to another protein from a species differing in their type of nucleic acid. The identification of characteristic domains using Pfam reveals that the majority of the identified proteins are associated with the replication machinery. To substantiate the results of the BLAST search, we conduct profile to sequence alignments and evaluate the conservation of motifs. Also, we conduct a comparison to proteins with resolved structure that exhibit the same characteristics.

By comparing the GC contents of the proteins to those of the corresponding viral species, we test whether horizontal gene transfer between the unrelated viruses can explain the observed homologies. The comparison however yields no indication of recent transfer events. For some proteins, a strong correlation of the GC contents to those of the host provides evidence for gene transfer between the virus and its host.

We hypothesise that the rates of evolution explain the observed similarities primarily between informational genes of unrelated viruses while we did not observe similarities between proteins forming the virus scaffold. Scaffold proteins rarely have relatives outside their immediate family. This might be attributed to a more rapid evolution of these proteins, making relatedness, if any, undetectable on sequence level.

The second analysis aims at the identification of viral proteins atypical for the virus family in which they are observed. We devise a prediction algorithm that employs a one class support vector machine to evaluate statistical signatures of viral genes. The algorithm is based on the assumption that nucleic acid sequences of a genome exhibit common statistical features specific to that genome, whereas genes recently transferred horizontally into the genome still have the original statistical features. Given all genes of a viral family, we use the genomic signatures and identify those genes that do not originate from any species in that family.

We evaluate the performance of the algorithm on real and artificial data sets. Artificial data sets are created by augmenting the set of genes of a family by random foreign genes. We show that the algorithm robustly recovers the artificial outliers, and thus can detect potential examples of horizontal gene transfer. By a combination of BLAST searches and literature research we conclude that the discovered atypical genes in the real data sets often coincide with literature evidence for horizontal gene transfer.
## Contents

1 Introduction .......................................................... 1

2 Biological Background ............................................... 3
   2.1 Life and Viruses ............................................... 3
   2.2 Hereditary Information ......................................... 4
   2.3 Molecular Properties of Proteins ............................. 6
   2.4 Taxonomy of Viruses .......................................... 6
   2.5 Evolutionary Mechanisms of Innovation ......................... 7
   2.6 Compositional Features of Genomes ......................... 8
      2.6.1 GC Content ............................................... 9
      2.6.2 Amino Acid Content ..................................... 9
      2.6.3 Codon Usage ............................................. 10
      2.6.4 Oligonucleotide Signatures ............................. 10

3 Related Study ....................................................... 11
   3.1 Methods to Detect Horizontal Gene Transfer .................. 11
   3.2 Studies on Gene Transfer in Viruses .......................... 12

4 Detection of Horizontal Gene Transfer Between Viruses .......... 15
   4.1 Motivation ...................................................... 15
   4.2 Approach ....................................................... 15
      4.2.1 Selection of Data ......................................... 16
      4.2.2 Finding Similar Proteins ................................. 16
      4.2.3 Identification of Protein Families ....................... 17
      4.2.4 Revealing Possible Gene Transfer Events ................ 18
   4.3 Results ......................................................... 19
      4.3.1 dUTPase Family ........................................... 20
      4.3.2 RNA Helicase Family ..................................... 23
      4.3.3 Phage Integrase Family .................................. 24
      4.3.4 Helicase C Family ......................................... 26
      4.3.5 Hema Esterase Family ..................................... 28
      4.3.6 RdRP I Family ............................................. 29
      4.3.7 Pkinase Tyr Family ....................................... 31
      4.3.8 OrfB IS605 Family ......................................... 33
      4.3.9 MMTV SAg Family ......................................... 34
      4.3.10 Parvo NS1 Family ......................................... 35
      4.3.11 GC Content Analysis ..................................... 36
   4.4 Discussion ...................................................... 38
Contents

5 Model to Detect Horizontal Gene Transfer Into Viruses
  5.1 Problem Setting ......................................................... 41
  5.2 Approach ................................................................. 42
    5.2.1 Prediction Algorithm ........................................... 42
    5.2.2 Preparation of Input Features .................................. 47
    5.2.3 Measures for Validation ........................................ 51
    5.2.4 Steps of Evaluation ............................................. 55
  5.3 Results ................................................................. 56
    5.3.1 Robustness of the Prediction ................................. 58
    5.3.2 Prediction Quality on Simulated Data ....................... 58
    5.3.3 Comparison of Ranking Potential on Real and Simulated Data 67
    5.3.4 Prediction Quality on Real Data ............................. 67
  5.4 Discussion .............................................................. 73

6 Conclusion .............................................................. 77

List of Figures ............................................................ 79

List of Tables .............................................................. 81

List of Acronyms .......................................................... 83

Bibliography ............................................................... 85
1 Introduction

Biology, as many sciences nowadays, has to deal with big data. Gigabytes of data from experiments are publicly available in databases. Its systematic analysis can help to broaden understanding of the mechanisms of life and evolution. In this thesis, horizontal gene transfer (HGT) in viruses is surveyed by computational analyses. HGT is the exchange of genetic material between organisms other than by reproduction and typically across species boundaries.

Viruses are a class of highly diverse organisms with origins and evolution still under debate. The reproduction of viruses involves takeover of a host cell. Cells under viral control alter the metabolism such that they produce all building blocks necessary to replicate the virus. The takeover, for some types of viruses, called retroviruses, implies the integration of their genetic material into the host genome. This retroviral mechanism is a typical and widespread example for HGT: If the host cell despite the virus occupation produces progeny, the integrated viral genetic information might be passed to the next generations and eventually become a permanent part of the genome of the host species. This gain of new genetic material is called lateral gene transfer (LGT) or HGT.

Because of their close interaction with the host cells, viruses are assumed to play a major role in evolution of life [Koonin et al., 2009]. Either viral genes might be left behind and incorporated in the host genome, as described above, or host genes are taken up by viruses and become integrated into the viral genome. Herein, particular focus is on the latter case where genes are taken up by viruses.

The term viruses comprises fairly complex organisms that exceed the size of some bacteria as well as minimalistic compositions of a few genes encapsulated by a protein. As a matter of fact, unlike cellular organisms, viruses are not classified according to their ancestry. They are grouped together because of their characteristic to infect other organisms and to make use of the foreign reproduction machinery to reproduce themselves. Where on one hand side a virus infection is a disease, that needs to be treated, another view is to regard viruses as promoters of evolution. By infecting the host, they exert selective pressure on the host to invent a defence mechanism. By evolving a defence strategy, the host then puts pressure on the virus to either adapt or to find a new host. When switching hosts, viruses are suspected to occasionally act as transmitters of genes between species [Forterre, 2006]. This means they might promote transfer of genes from one species to another and thereby act as a driving force in evolution. There is evidence for both the presence of viral genetic material in non-viral genomes [Liu et al., 2012], as well as acquisition of host genes by viruses [Baldo and McClure, 1999]. In this study cases of gene adoption by viruses are systematically identified.

To date, HGT into and between viral species has not been systematically studied. However, for individual viral species, gene transfer has been analysed. Herpesviruses for instance were shown to have attained genes for dUTPase by HGT [Baldo and McClure, 1999]. Another well-described example is Mimivirus and its phage Sputnik, that have been shown to share homologous proteins [La Scola et al., 2008].

Two different aspects are the focus of this study: The first is to investigate whether there exists evidence for transfer between viruses. We aim to find out if viruses not in relation to
one another have acquired similar genes; either because genes were obtained from a common source, or because there exists a not yet discovered mechanism of gene exchange among viruses, or because of undiscovered ancestral relations between some classes of viruses. The second part of this thesis poses the question whether there exist viral genes that can be suspected to have alien origin. We devise a model to systematically predict which viral genes do not originate from the species they were found in. These genes presumably have been integrated into virus genomes only recently. As besides a few anecdotal studies there is no gold standard dataset of viral genes acquired via HGT, validation of the model is particularly demanding. Finally, selected results from both analyses are examined for their integrity.

The data employed in this study comprises all genetic sequences of viral origin available in European Nucleotide Archive (ENA) [Leinonen et al., 2011] as well as the corresponding protein sequences from UniProt Knowledgebase (UniProtKB) [Magrane and Consortium, 2011].

Structure of this Work

The thesis has the following structure: First, the reader is introduced to the field of study. In chapter 2 an insight on the biological background about gene transfer and its role in evolution are provided. Also measures for statistical assessment of compositional features of genomes are introduced. Next, in chapter 3, related work on identification of gene transfer events is described. Particular focus is on studies that involve viruses. The following two chapters approach the methodology. In chapter 4, the identification of homologous proteins in unrelated viruses is described. These proteins with common origin potentially have been acquired by gene transfer events. The identification is accomplished by means of sequence comparisons and analysis of statistical features. In chapter 5, the related question of finding viral genes with potential foreign origin is addressed. For the identification of genes atypical for the viral family they are observed in, a one-class support vector machine (SVM) algorithm is introduced and evaluated. In each of the two chapters, the methodology and employed tools and data are introduced first. Next, the results are provided and discussed. Finally, chapter 6 offers final conclusions and an outlook.
2 Biological Background

This chapter provides a brief overview of biological foundations necessary to understand mechanisms of evolution. The focus is particularly on HGT and on the role viruses play for the evolution of life. Furthermore, properties of genomes are discussed which allow for statistical assessment in order to distinguish organisms.

2.1 Life and Viruses

Living beings are highly diverse yet have a common ground. They are made of cells and share the basic machinery to reproduce themselves. Each species is able to produce progeny of the same kind imparting the hereditary information to the offspring. This concept of hereditary is primal for the definition of live [Alberts et al., 2008].

Traditionally, the living things are classified by their outward appearances and since recently also by analysing similarities in their hereditary information. As Darwin claimed, close resemblances between organisms can be interpreted in terms of evolution from common ancestors [Darwin, 1859]. These ancestry relations define a tree of life for all cellular organisms that has three major divisions: Bacteria, Archaea, and Eukaryota.

Viruses belong to none of these three domains. They are particles that carry their hereditary information in the form of nucleic acid and are able to reproduce themselves only by invading another cell that is usually called host cell. At its simplest, a virus consists of nucleic acid surrounded by a protective protein coat. In some viruses, this nucleoprotein is surrounded by further proteins or a lipid bilayer. With the outermost proteins, viruses recognize the correct host and gain entry to the cytoplasm [Dimmock et al., 2009]. For their reproduction, viruses have to employ other organisms as they do not have full metabolism and a reproduction machinery. The definition of viruses originates back to a time where there did not exist methods to reveal these infectious particles in experiments. Only the observation that there is something infectious to organisms led to the definition. So, despite the great variety of viruses they are all classified together. And for this highly diverse class, no unifying link to the tree of life has been established yet.

Viruses are ubiquitous. They are found in almost every ecosystem on earth and infect all types of organisms including animals, plants, bacteria, and archaea. There are approximately $10^{31}$ viruses on Earth [Breitbart and Rohwer, 2005]. Initially, viruses were discovered by Dimitri Ivanovskv in 1892 and by Martinus Beijerinck in 1898 who studied the tobacco mosaic disease and could show that infected plant sap remained infectious even after filtering through finest filters that would catch all types of bacteria [Lecoq, 2001]. Since this discovery, more than 5,000 viral species have been described [Dimmock et al., 2009].

Whether and how viruses are related to other living organisms is actively discussed. There are several hypotheses about the origin of viruses. The escape hypothesis states that viruses arose from mobile genetic elements that gained the ability to move between cells. The reduction hypothesis claims that viruses are residuals of cellular organisms; highly specialized parasites
that over time lost all not required genes. The virus-first hypothesis states that viruses existed in a pre-cellular world as self-replicating units and then co-evolved with their hosts. Possibly each of these theories has its point, perhaps only for a particular portion of viruses and not for all of them at a time. For each hypothesis, there exist examples among the viruses that support the argumentation. It is likely that viruses have arisen multiple times independently rather than from a single progenitor [Dimmock et al., 2009].

Aside from the discussion whether viruses are a form of life, there exist multiple pieces of evidence that viruses exhibit evolution. Their hereditary information is similar to those of cellular organisms, and for reproduction they even use cellular organisms. Although the overall picture of relatedness of viruses is unclear, for single virus families, evolutionary trees are reasonable.

In recent literature [Raoult and Forterre, 2008, Philippe et al., 2013] it is debated whether some viruses, rather large and complex ones, should be placed in the tree of life as they exhibit many similarities to living organisms and might form a fourth major division besides Archaea, Bacteria and Eukaryota. Other viruses, on the contrary, are so simple that doubts about regarding them as a life form appear to be justified.

2.2 Hereditary Information

All living cells store their hereditary information the same way. It is encoded in a linear molecule called deoxyribonucleic acid (DNA). These molecules are long linear chains that form in most cases double helices bound by hydrogen bonds and always consist of the same monomers. Each monomer consists of a sugar (deoxyribose in the case of DNA) and a base: adenine (A), cytosine (C), thymine (T), or guanine (G). While the sugars covalently linked by phosphates produce the chain backbone, the bases make hydrogen bonds to the pairing strand. Both strands together exhibit a helical structure, the DNA double helix. The pairing strand always is complementary as A only binds to T and G only binds to C. Strand complementarity makes half-conservative replication of DNA possible. It is never synthesized freely in isolation but always along a template of already existing DNA [Alberts et al., 2008].

The hereditary information contains templates for all proteins necessary to construct the organism. Triplets of three bases, also called codons, encode for specific amino acids. When the DNA is read in triplets from a defined starting position to a stopping position, the information for a sequence of amino acids results. The range from starting to stopping position is the coding region of a gene. A protein coding gene refers to a unit of information consisting of the coding region and all pieces of DNA necessary to regulate the transcription of the coding region. The resulting amino acid sequence forms a protein, that folds properly spontaneously or with the help of the intracellular machinery. The genetic code describes which triplet yields which amino acid. It is used by all forms of life universally, except for a few minor variations. An overview which codons yield which amino acids is provided in figure 2.1.

In a sequence of DNA, coding regions are often not directly subsequent one another, but are interspersed with non-coding regions. While in eukaryotes the amount of non-coding regions is immense, for example over 98% in the human genome [Elgar and Vavouri, 2008], viral hereditary information is typically highly compact and mechanisms exist to further increase the information content. The information for several proteins can be encoded in concatenation with one start codon prior to the first gene and one stop codon subsequent to the last. These genes require an additional proteolytic step to obtain the final proteins as the whole information is translated at once into a single polyprotein. Another possibility is the presence of multiple over-
2.2 Hereditary Information

Figure 2.1: The universal genetic code. Read from the inside to the outside, this illustration shows which triplets of nucleotides encode which amino acids. Source: http://www.chemgapedia.de/vsengine/media/vsc/de/ch/5/bc/gen_protein/bilder/gensonne.jpg

lapping reading frames. The same nucleotides read from a shifted starting position can yield a different protein. Alternative splicing, a mechanism also known from eukaryotes, may yield different variants of a protein by excluding or altering parts of the gene after translation. Due to the small amount of non-coding hereditary information in viruses, we solely focus on coding sequences.

Hereditary information of some viruses is not stored in the form of DNA but in ribonucleic acid (RNA). RNA is similar to DNA but uses uracil (U) instead of T, and ribose instead of deoxyribose as the sugar. As suggested by the alternative name for T, 5-methyl-uracil, T and U are similar: T is methylated U. All organisms use RNA as an intermediate template for the synthesis of proteins. Only particular viruses employ RNA as the permanent storage for their hereditary information. Cellular organisms always use the more stable DNA to store hereditary information. For notational convenience, we will neglect this fact when it comes to comparisons between sequences of hereditary information and treat all as if it was DNA, substituting all Us with Ts.

The entire hereditary information of one species is called genome. Similarly, on the level of proteins, the complete proteome describes the set of all proteins that belong to one species. It is called complete because the term proteome is already defined as the entire set of proteins expressed at a time.


2 Biological Background

2.3 Molecular Properties of Proteins

Proteins are the building blocks of life. They are chains of amino acids with a particular spacial structure, their characteristic fold. Proteins are synthesized by organisms according to a genetic template. The folds of proteins are highly conserved throughout evolution. The local three-dimensional structures, called secondary structure elements, that occur within proteins are alpha helices and beta strands. These elements are connected by sequences of flexible structure, called turns. On the basis of these blocks super-secondary structure elements are defined. These are for instance barrels, beta sheets, helix-turn-helix and some other common motifs. Super-secondary structures are combinations of secondary structures which occur in many proteins irrespective of their origin. The tertiary structure is the complete spatial structure of the protein. It is a combination of secondary structural features. Structural units that fold independently and often bear a specific function are called domains. Domains are highly preserved during evolution.

Only few amino acid residues directly act in the process of protein performing its function (e.g catalysis in enzymes), and they are called protein functional or catalytic site. Mutations in functional sites mostly have negative effect on the functionality which may result in lethality of the whole organism and hence no progeny from organisms with such mutations.

Protein structure is generally more conserved than sequence throughout evolution. Genetic sequences and protein sequences may vary but for an organism to work, the protein function is critical and the function is directly connected to the protein fold. In particular, for the study of virus relations, due to high mutation rates of some viruses, the analysis of folds might bear more fruit than the analysis of sequence. Viruses evolve very quickly and exhibit a high mutation rate. The Human immunodeficiency virus (HIV) for instance exhibits an extraordinary scale of variation [Korber et al., 2001] showing levels of nucleotide sequence similarity of 50% or lower [Vanden Haesevelde et al., 1994] between different strains of the same virus.

2.4 Taxonomy of Viruses

A taxonomy is the classification, identification, and naming of organisms. Historically, taxonomies have been set up upon morphological comparisons. Nowadays, devising a taxonomy implies to study the evolutionary relationships among groups of organisms using molecular sequencing data. The taxonomy is then established on the basis of phylogenetic reconstructions. A hypothesis about the order of evolutionary events can be visualized by means of a phylogenetic tree. It depicts a branching process of how species alter over time [Haeckel, 1866].

To establish a phylogeny, comparisons of organisms are carried out. These comparisons can be on the level of genes or proteins. Genetic sequences are compared by means of alignments, or through alignment-free methods exploiting statistical properties of the sequences. On the level of proteins, amino acid sequences and protein structures are compared. In all cases, the goal is to identify structural, functional, or evolutionary similarity between sequences assuming that similarity in sequences reflects such higher level similarity. If similar genes are actually evolutionary related, they are called homologous to each other.

While cellular organisms can be placed into the tree of life, viruses are primarily classified according to their phenotype. The distinguishing properties are their morphology, their nucleic acid type, their mode of replication, the host organisms, and the type of disease they cause. For the lack of evidence for a common origin, they are not placed in the tree of life but are named by an independent system. The system by the International Committee on Taxonomy of Viruses
2.5 Evolutionary Mechanisms of Innovation

Evolution is the establishment of innovation in the hereditary information. By preservation and accumulation of innovation over generations, different species evolve from a common ancestor. This idea might seem to contrast to the above description of how DNA is copied and inherited to the offspring: Progeny should be identical to the parent. However, the copying mechanism is not perfect. Random copying errors cause alternation of the nucleotide sequence that is handed to the offspring. Such mutations may either have a positive effect, no effect, or a negative effect. Where in the first case, the resulting organism has an increased chance to reproduce itself, the
latter case yields to a decreased reproduction rate, sometimes down to no progeny at all. Mutations in conserved genes often lead to deleterious effects, and these genes stay almost unchanged over a tremendous amount of generations. Whether a mutation is of advantage, sometimes depends on the environment of the organism. The adaptation to varying environmental conditions can be regarded as the driving force of evolution.

Copying errors within a gene are called intragenic mutations. Other mechanisms how innovation can be introduced to a genome are gene duplication, segment shuffling, and horizontal gene transfer (HGT). Gene duplication initially yields two identical genes within the same cell. If they are both passed to the offspring over generations, they may diverge through copying errors and eventually even attain different function. Such genes are called paralogous to each other. Segment shuffling describes scenario of genes being broken and rejoined in an altered order. Horizontal transfer refers to the event of a piece of DNA being transferred from the genome of one cell to that of another. This transfer can even occur across species. It is called ‘horizontal’ or ‘lateral’ transfer to contrast from the usual process of transferring information from the parent to the progeny which is sometimes called ‘vertical’ [Alberts et al., 2008]

In particular for bacterial and archaeal evolution, horizontal transfer of genes is an essential mechanism [Alberts et al., 2008]. The ability to adapt to new environments often is attained by uptake of a new gene rather than by mutation of an existing one. Bacteria have acquired several mechanisms to horizontally exchange genetic material: Transformation describes the uptake of free genetic material from the environment and incorporation into the genome. Conjugation requires cell-to-cell contact: via a connection between two cells, even long fragments of DNA can be transferred. Another mechanism is transduction. This mechanism requires a phage, a bacteria-infecting virus, to take up genetic material and to donate it. The donated material can be either of viral or of bacterial origin. As the phage needs to bind on the cell surface, this mechanism is primarily common among closely related bacteria, whereas the other two also work for distant species. [Rakhuba et al., 2010]

2.6 Compositional Features of Genomes

Evolution takes place at different rates, yet it is slow compared to time spans we can comprehend. For this reason, to study the paths of evolution, it is necessary to deduce the order of innovation events over time from one single snapshot of the world today. Analysis of compositional features of a genome can identify characteristics which permit to distinguish organisms according to their evolutionary history. Such features are often called genomic signatures. Genomic signatures can be generated from an arbitrary fragment of a genome yielding a species specific result. This result is a characteristic vector of numbers\(^1\).

A good computational genomic signature fulfils two properties, pervasiveness and species specificity. Given random fragments of a genome with sufficient length, pervasiveness assures their signals to be the same (or at least similar). Species specificity grants different signals for different species. Thus, if a gene exhibits a genetic signature different from the rest of the genome it is embedded in, it is likely to be recently laterally transferred. The characteristic signatures originate in the characteristic evolutionary processes each species undergoes in time. The genome composition for each species is determined by selection and mutational pressure [Sueoka, 1988]. Comparison of compositional patterns enables the identification of

\(^1\)Genomic signature is not used consistently throughout literature. In other contexts, the term describes a unique short sequence of nucleotides identifying an organism.
horizontally transferred genes as those whose features are atypical for a particular genome. However, only recently acquired genes would be detected because sequences quickly adjust to their new genome pattern [Daubin et al., 2003].

In the remainder of this section, different genomic signatures are introduced. All of them capture global features of the sequences. Because of the species specificity and pervasiveness of genome signatures, these signatures can be used to characterize and identify a genome on the basis of a short genome fragment from that source. For all these genomic signatures to work well, the sequences must be sufficiently long. Very short sequences exhibit large biases especially if the dimensionality of the signature is large. Genomes of viruses however can be very short inducing limitations to the applicability of genomic signatures.

### 2.6.1 GC Content

GC content is an early discovered compositional feature [Sueoka, 1963]. It is a popular characterization for genomes in lines with the genome signature definition. GC content measures the ratio of G plus C in the nucleotide sequence. Across the tree of life, the GC content ranges from 16.5% in *Carsonella ruddii* to 75% *Anaeromyxobacter dehalogens* [Nalbantoglu, 2011]. Its variation has been attributed to varying environmental parameters such as growth temperatures. Ecological constraints determine preferences in the ratio of G and C over A and T. Cytosine and guanine form three hydrogen bonds between the strands of the DNA double helix, providing more stability and resistance towards denaturation. Adenine and thymine form only two hydrogen bonds but their production costs are lower [Saenger, 1984].

Pattern in usage of the different bases in DNA initially have been described by Chargaff and colleagues [Zamenhof et al., 1950, Rudner et al., 1968]. Chargaff’s first parity rule states that for double stranded DNA the proportion of A equals that of T and the proportion of C equals that of G. The second parity rule states that proportions are maintained not only globally but in each of the two strands. Chargaff also observed that the composition of DNA varies from one species to another. Mitchell and colleagues [Mitchell and Bridge, 2006] showed that the rule is not applicable to ssDNA genomes or any type of RNA genomes. Despite these limitations to Chargaff’s rule we may use the GC content to characterize virus species. For them however, except for dsDNA viruses, it is not possible to infer the proportions of all four nucleic bases from the ratio of G plus C.

A signature closely related to the GC content is the third position GC content (GC3s). For this measure, only the nucleotides in the third position of every codon are taken into account, because for many amino acids the synonymous codons are degenerate in the third position, and hence this position is not under evolutionary pressure.

### 2.6.2 Amino Acid Content

The amino acid content represents the relative frequencies of amino acids used in a protein as a 20-dimensional vector. Analogous to GC content at the genome level, certain organisms have a tendency to specific amino acid prevalences in their proteins. The result is a spectrum of typical amino acid usage pattern among various taxa. According to Nalbantoglu [Nalbantoglu, 2011], it has been suggested that the species specificity of amino acid usage is the outcome of certain evolutionary processes, such as the response to different environmental temperatures, the economy of nutrient supply, and the susceptibility to oxidation.
2 Biological Background

2.6.3 Codon Usage
Codon usage is a vector describing how often each triplet in a genetic sequence is used. Codons are the triplets of nucleotides which uniquely encode for an amino acid. The same amino acid can be encoded by several synonymous codons. The translation from codons to amino acids is depicted in figure 2.1. The codon usage can be either described as absolute proportions of occurrences of every possible codon, or by means of relative amounts of the codons encoding for the same amino acid.

2.6.4 Oligonucleotide Signatures
Oligonucleotides are short nucleotide sequences, typically no longer than 15 residues. The oligonucleotide signature is a vector of the relative frequencies of all possible oligonucleotides with the defined length throughout a sequence. While the previously described signatures accounted for biological properties, oligonucleotide frequencies have no specific biological meaning. Regarding a gene purely as a sequence of letters from the alphabet \( \Omega = \{A, T, C, G\} \), a sliding window of a certain length \( l \) is moved along the sequence. All words observed in that window are counted yielding a \(|\Omega|^l\)-dimensional vector of frequencies of all possible words. Oligonucleotides are also referred to as k-mers or n-grams.

Relative abundances of dinucleotides, oligonucleotides of length two, have been originally described by Josse and co-workers [Josse et al., 1961] and have been studied extensively by Karlin and colleagues [Karlin and Burge, 1995, Karlin, 1998]. They observed characteristic pattern preserved throughout genomes. Dinucleotide signatures were also found to have the tendency to be more similar among closely related organisms than among distant ones. Pride et al. [Pride et al., 2003] extended the concept of from dinucleotides to oligonucleotides. By consideration of longer words, and with this a much higher number of parameters, it is more likely to capture deviation from the genome background compositional distribution of the genome. This requires enough data to produce reliable probability estimates. In short or biased sequences however, a significant amount of data is likely to be missing.
3 Related Study

In this chapter, an overview of studies about HGT is given. Focussing on methodological differences, we first provide a review of different approaches to detect evidence for HGT. Subsequently, we provide examples of HGT studies addressing viruses.

3.1 Methods to Detect Horizontal Gene Transfer

Although studying HGT in viruses is hampered by their high evolutionary rates, short genetic sequences, and the lack of a common ancestry, it is inspiring to survey how HGT is analysed in other organisms.

Discussing HGT in prokaryotes, Koonin and co-workers present a collection of approaches for the detection of HGT [Koonin et al., 2001]: One option is to observe an unexpected ranking of sequence similarity among homologs. This means, a sequence from a particular organism shows the strongest similarity to an evolutionary related sequence from a taxon which is not closely related to that organism. Another possibility is to come across an unexpected phylogenetic tree topology. The analysis of phylogenetic tree topologies is traditionally the principal means to decipher evolutionary scenarios. However, as Koonin et al. point out, phylogenetic analysis is time and labour consuming, critically depends on correct sequence alignments, and is hard to automate. Another option, with many complete genome sequences available, is the study of unusual phylogenetic patterns. Sets of organisms related by vertical descent, called sets of orthologs, are the basis for this study. In sets of organisms related by inheritance the presence of a gene only in one lineage might indicate HGT, in particular if that gene is typical for organisms unrelated to those in the sets of orthologs. A further method is the study of conservation of gene order between distant taxa. This method is predominantly applicable in bacteria that maintain operon structure. Operons, collection of genes under the control of a single regulatory signal, are maintained during evolution. With the observation that in distant genomes it is unlikely to find three or more genes in the same order unless these genes form an operon, HGT seems to be the most likely scenario when a (predicted) operon is present in only a few distantly related genomes. To detect HGT, also it is an option to discover genes with anomalous nucleotide composition. Although widely used, this method is applicable only to discover recent gene transfers. No matter which approach is used, harder than detecting that gene transfer occurred is to determine which organism is the donor and which one is the recipient. Furthermore, all indications for horizontal transfer necessarily remain probabilistic because there is no record of the evolutionary events [Koonin et al., 2001].

In other studies, genetic material of *Bacillus subtilis* has been subjected of HGT identification using hidden Markov models (HMMs) [Bize et al., 1999, Nicolas et al., 2002]. These statistical models are capable to identify segments of DNA with composition of oligonucleotides different from the species background. Bize and co-workers could identify eight DNA segments, potentially issued from HGT. Most of the genes involved in these segments were of unknown function. Two of the genes however could be identified as tetracycline resistance genes. Nico-
3 Related Study

Las et al. refined these results noticing that the regions of atypical nucleotide composition are AT-richer than the host genome raising the question about their origins.

Contrasting Koonin’s doubts on the automatability of phylogenetic tree analysis, Boc and Makarenkov introduce a computational method for the detection of unexpected topologies in phylogenetic trees [Boc and Makarenkov, 2003]. The method is based on the computation of differences between topologies of species and gene phylogenetic trees. It applies a least squares optimization to test the possibility of HGT between any couple of branches of the species tree.

Working with sets of orthologs, Suchard presents a stochastic algorithm to identify potential events of HGT taking a random walk over the space of phylogenetic trees. That walk allows for the development of a joint probabilistic distribution over multiple gene trees and an estimable species tree [Suchard, 2005].

Also SVMs have been employed to detect HGT. Tsirigos and Rigoutsos describe a one class SVM approach that uses a sliding window to identify regions of atypical composition in genomes [Tsirigos and Rigoutsos, 2005b]. The feature vectors which serve as input to the SVM are constructed on the basis of compositional features such as higher order nucleotide sequences. Their construction is subject of a preceding publication [Tsirigos and Rigoutsos, 2005a].

The identification of horizontally transferred genes confirms the existence of HGT mechanisms, but to understand the role of these mechanisms in evolution, identification is not enough. Galtier presents a model of genome evolution with HGT. This model is used to simulate sequence data from multiple genes with our without HGT in order to study the consequences of HGT for phylogenies. It should provide quantitative assessment of the impact of HGT on phylogenies and of the ability of tree building methods to cope with such events [Galtier, 2007]. This approach is substantially different to all mentioned before. It aims at understanding the underlying mechanism and not at describing the results this mechanism produces. Also, it helps to assess the robustness of tree construction methods under the occurrence of HGT.

3.2 Studies on Gene Transfer in Viruses

In the previous section, different methods for the identification of horizontally transferred genes have been introduced. Most of these methods aim at cellular organisms and would be applicable to viruses only to a limited extend, if at all. There is no common ancestor and hence no unifying phylogenetic tree for viruses. Sets of orthologs can be constructed only for closely related viral species. It has also been shown that, despite a rapid increase of the number of sequences of viral genomes, there are much more viral genes yet to be discovered [Kristensen et al., 2013]. Identifying HGT by means of statistical analysis appears to be the only suitable option that involves however additional challenges for virus genomes: Statistics are likely to be biased due to the comparably short viral genomes, and viruses tend to exhibit high mutation rates. Despite these difficulties, a number of publications approach the analysis of HGT in viruses.

The world of ocean microbes is rich and dynamic and is a source for major discoveries on the interaction of viruses with their microbial hosts. Interaction between phytoplankton and viruses is the subject of various studies. Sullivan and co-workers focus on cyanobacterial viruses and their hosts. They could show that whole genes of the photosystem reaction center have been transferred from host to phage [Sullivan et al., 2006]. The host-like photosynthesis genes of the cyanophages presumably augment the host photosynthetic machinery during infection. Their study is experimental and involves screening of cultured cyanophages of known family and host range and viral DNA from field samples.
3.2 Studies on Gene Transfer in Viruses

Eukaryotic algae, as well found in phytoplankton, and their interaction with viruses are subject of a study by Monier and co-workers. They report evidence for the transfer of multiple genes involved in the sphingolipid biosynthesis pathway between the eukaryotic microalga *Emiliania huxleyi* and its large dsDNA virus *Emiliania huxleyi virus* [Monier et al., 2009]. The study involves gene sequencing and a phylogenetic analysis. It does not reveal the direction of the HGT. If the transferred genes are of viral origin, ancient viruses must have controlled the complex metabolic pathway in order to infect primitive eukaryotic cells. If the genes originate from the host, the serial acquisition of genes involved in the same metabolic pathway might have been an advantage for viruses in the evolutionary race for survival.

Studying organisms acidic and alkaline hot springs, Schoenfeld and colleagues find evidence for HGT of family A DNA polymerase between thermophilic viruses, and the bacterial phyla of Aquificae and Apicomplexa. They attribute the appearance of the species nowadays not only to one event of gene transfer but to a whole network of transfers having taken place during the time course of evolution [Schoenfeld et al., 2013].

The family of *Poxviridae* has been subject of a HGT study which revealed that proteins encoded by members of the subfamily *Chordopoxvirinae* exhibit greater similarity to eukaryote proteins than to proteins of other virus families illustrating the important role of gene capture from the host for virus evolution [Odom et al., 2009].

A special case is the story of *Sputnik virus*. This virus infects *Acanthamoeba polyphaga mimivirus*, one of the largest known viruses. Some of the genes of *Sputnik virus* apparently originate from *Mimivirus* [La Scola et al., 2008]. *Sputnik virus* is the first virus discovered that uses the protein machinery of another virus to propagate. This discovery opened up a broad discussion: It is suggested that the virophage could perform HGT between viruses, analogous to the way that bacteriophages transfer genes between bacteria [Pearson, 2008]. Also, the question was raised whether *Mimivirus* is actually a virus. With its enormous size compared to other viruses, it already has a unique position within the virus world. The isolation of the virophage Sputnik revealed that *Mimivirus* is susceptible for virus infections, very much alike those of cellular organisms [Claverie and Abergel, 2009]. In addition, *Mimivirus* encodes for genes that were never encountered in viruses before. This yields to the hypothesis that giant viruses coexisted with the cellular ancestors and represent a distinct but related group along with the kingdoms Archaea, Bacteria, and Eukaryota [Nasir et al., 2012]. The recent discovery of *Pandoravirus*, an amoeba virus larger than any other characterized virus, seems to strengthen the hypothesis about at least some viruses constituting a fourth domain of life. More than 93% of *Pandoravirus* genes resemble no known genes and their DNA polymerase clusters with those of other giant DNA viruses [Philippe et al., 2013].

All these studies focus on individual viral species or at most one family revealing interesting details. In contrast, we aim to provide systematic assessment of HGT in all known viruses at once. By not focusing on a particular virus, we might not come across results as revealed in detailed work. But we can hope to contribute to the broader picture of the role of HGT in viruses.
4 Detection of Horizontal Gene Transfer Between Viruses

In this chapter, we study similarities between proteins of unrelated viruses. It involves comparisons of protein sequences and protein structures of viruses differing in the type of nucleic acid they use to store the genome. Similarities are primarily observed among proteins related to the reproduction machinery. The observed resemblances are discussed in detail for each found family of proteins. In particular, we analyse whether an evolutionary relationship between the found proteins is likely. Part of this analysis is the survey of GC contents of the corresponding genes and organisms.

4.1 Motivation

Virus evolution is a field of ongoing research. Despite advances in uncovering evolutionary relationships among organisms, the origin of viruses is still under debate. As discussed in chapter 2, different hypotheses about the origin of viruses exist and it appears plausible that viruses do not have one single common evolutionary origin. In this chapter, we approach the question whether viruses despite controversies about their origins have any common ground. We are not seeking for a unifying ancestor that probably does not exist. But we investigate whether viruses distant in classification, with different replication strategies, have related proteins. Such proteins might originate from HGT either between the viruses or from a host common to both viruses, or from two different hosts that have harboured related proteins. We restrict the analysis to comparing viruses from different major groups of the Baltimore classification which classifies viruses by the nucleic acid type of their genome. Thus, viruses from different groups have different types of nucleic acid, and so are very distantly, if at all, related. Relations on this level are impossible to find by phylogenetic methods as they focus on inherited properties. Phylogenetic analysis is applicable on the level of virus families where the relatedness of species is clearly traceable by descending from a common ancestor. On the level of different Baltimore classes, we conduct a protein similarity search of which we report and interpret the findings. Clearly, this will not allow to construct a hypothesis about relatedness of different virus types. However, it will reveal whether there is any (recent) genetic exchange between viruses with different genetic material.

4.2 Approach

Candidates for examples of horizontally transferred genes are obtained by means of a stepwise procedure. Starting with all viral complete proteome sequences, we employ different tools and methods. After preliminary filtering steps, detailed analysis of the candidates is carried out using multiple sequence alignments and assessing GC contents. The individual steps of the acquisition are described subsequently.
4 Detection of Horizontal Gene Transfer Between Viruses

4.2.1 Selection of Data

Detailed comparison considering structural features of all viral genes is infeasible. Therefore, an appropriate subset of the data available has to be selected. The choice of the data set has been made trading off as much data as possible against reasonable computational effort and as little junk data as possible. Biological databases grow rapidly today and might, despite good intentions, contain incomplete records. For instance metagenomics, the analysis of genetic material recovered from environmental samples, contributes a lot of information to databases but typically does not provide complete genomes of a species, often not even complete genes.

The dataset chosen for this study is the set of complete proteomes of viruses. It has been acquired from the UniProtKB [Magrane and Consortium, 2011] in September 2012 and contains 25,317 proteins from both Swiss-Prot and TrEMBL. This set has the advantage of containing only complete proteins. As each organism is represented in this set only once, also the redundancies in the data are kept at a minimum.

UniProt Knowledgebase The UniProtKB is the major database which is provided by the UniProt [Consortium, 2013]. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). It provides a resource for protein sequence and annotation data. Most data in UniProtKB, 98%, is derived from nucleotide sequence resources by the International Nucleotide Sequence Database Collaboration (INSDC) via automated translation of coding sequences (CDS). CDS themselves are either predicted or experimentally verified. UniProtKB comprises two sections. One is called Swiss-Prot and contains manually annotated records with information extracted from literature. The other section, TrEMBL, contains computationally analysed records waiting for manual annotation.

4.2.2 Finding Similar Proteins

Similar proteins in the set of complete proteomes are obtained conducting a sequence similarity search with the Basic Local Alignment Search Tool (BLAST). In the BLAST search, each protein sequence is submitted once as query to search the database of all other sequences. The algorithm returns a list of similar proteins for each query sequence. The list is equipped with measures indicating the amount of similarity of each result entry to the query. Due to the extensive length, the search results are pruned on basis of statistical criteria before further evaluated. A hit between two sequences is considered further only if it has an alignment length of more than 50 residues, a sequence identity of more than 30%, and an expect value of less than $10^{-5}$. The expect value describes the number of hits expected to be seen by chance with the observed score or higher in a random database of the same size.

We are interested in only a portion of the BLAST results, namely those between proteins from unrelated viruses. By unrelated viruses we mean those distinguishable by the type of nucleic acid according to the Baltimore classification introduced in section 2.4. Evaluating the taxonomy information of each protein, in total 160 hits are obtained between proteins of unrelated viruses. These hits are potential examples for HGT in viruses and further analysis is carried out on them in order to find evidence whether a common origin of these proteins is likely.
4.2 Approach

**Basic Local Alignment Search Tool**  BLAST [Altschul et al., 1990] is an algorithm for comparing biological sequence information. It is applicable to both amino acid sequences of proteins and nucleotides of genetic sequences. The purpose is to find regions of local similarity between sequences. BLAST compares the sequences against sequence databases and provides measures of statistical significance for each match. It was originally developed in 1990 and today is frequently used to search in databases. The popularity of BLAST issues from the considerable runtime advantage compared to classical dynamic programming approaches for sequence alignments like the Needleman-Wunsch or Smith-Waterman algorithm. This advantage is attained by means of a heuristic procedure.

The algorithm tries to limit the search space in order to decrease runtime. In contrast to the naïve approach to compare every residue against every other, BLAST uses short word segments to create alignment seeds. After an alignment is seeded, it becomes extended according to a user-defined threshold. A cut-off score is used to select the sequences that share significant similarities. If a hit is found, the algorithm checks whether a word with a sufficiently high cut-off score is contained within a longer aligned segment pair. The alignment terminates if an alignment score starts to decrease below a certain threshold.

### 4.2.3 Identification of Protein Families

Proteins are composed of functional regions, called domains. These independently folding units form the building blocks for different proteins allowing for the identification of proteins based on the combinations of domains. Also, insights into the function of proteins can be attained from the analysis of domains. The Pfam database [Punta et al., 2012] offers a large collection of protein domain families, where families are sets of protein regions that share a significant degree of sequence similarity. Each family is represented by multiple sequence alignments and HMMs.

The first step of the analysis of the BLAST search results is to check if the protein families for each hit are consistent. For meaningful hits we expect the protein pairs to show similar architectural features. The Pfam database provides a tool to identify the protein family given a protein sequence. Using this tool, we compare if the proteins match.

Manual comparison to the seed alignments from the Pfam database for each protein family observed in the BLAST search result further substantiates the findings. Seed alignments are sequence alignments that contain a small set of representative members of the family. Specifically, we inspect the active sites of the proteins for their conservation of properties. Where available, the comparison also involves analysis of a family representative with resolved 3D structure. Such structures are obtained from the Protein Data Bank (PDB) [Berman et al., 2000].

The proteins from the result set as well as the representative 3D structures are aligned to the profile HMM derived from the seed alignment of the respective family. While the seed alignments are plain multiple sequence alignments, profile HMMs are statistical models of a protein. They carry information how likely each amino acid is in each position and are implemented by means of HMMs using states and transition and emission probabilities between them. In other words, profile HMMs provide a position-specific scoring system for substitutions, insertions, and deletions. The alignment to the profile HMMs is carried out using HMMer [Finn et al., 2011]. HMMer offers a dynamic programming algorithm that finds the most probable path that the sequence may take through the profile HMM and reports the score attained with this path.
All alignments presented in this work are visualized using TeXshade [Beitz, 2000]. The 3D structure representations of proteins were generated using the molecular graphics program PyMOL [Schrödinger, 2010].

**Pfam Database**  The Pfam database [Punta et al., 2012] is a database of protein domain families provided by the Wellcome Trust Sanger Institute. Entries in the Pfam database comprise a protein sequence alignment and a HMM, a statistical model to represent the protein family. If available, the entries are supplemented with the information on experimentally resolved 3D structures of representatives of the family. With this information, the database allows for exploration of structure and domain architecture. In addition, it provides higher level groupings of related protein families, so-called clans. Pfam consists of two parts, Pfam-A and Pfam-B. While the information in Pfam-A is manually curated, Pfam-B contains automatically curated entries waiting for manual inspection. One particular feature of Pfam-A is the availability of seed alignments for each family. Seed alignments are multiple sequence alignments of a small set of representative family members. From these alignments, statistical models can be derived, profile HMMs.

**Protein Data Bank**  PDB [Berman et al., 2000] is a repository of information about the three dimensional structures of large biological molecules. Protein structures predominantly obtained by X-ray crystallography or nuclear magnetic resonance spectroscopy are submitted to this resource by biologists and biochemists from around the world. The requirement of many journals for scientists to submit their structures to this database as well as the manual reviewing process for newly added structures makes PDB a major resource of structural information.

**HMMer**  HMMer is a software package for searching sequence databases for homologs of protein sequences, and for making protein sequence alignments. It can compare profile HMMs to either a single sequence or a database of sequences.

### 4.2.4 Revealing Possible Gene Transfer Events

To find evidence whether HGT could explain the result of the conducted BLAST search, compositional features of the coding sequences of the proteins are taken into account. The GC content, as introduced in section 2.6, is a species specific signature that is easy to obtain. We check for discrepancies between the characteristic GC content of a species and the GC content in the gene of interest from that species. A large deviation is usually assumed to indicate that a gene might have been acquired through HGT [Koonin et al., 2001].

In a coding sequence, we measure the GC content in the third position of each codon only. This position is often ambiguous among synonymous codons, and thus the strands under no evolutionary pressure in these cases. So this position is most likely to retain the GC pattern characteristic to the donor species. The analysis of the third position GC content (GC3s) is carried out with the help of CodonW [Peden, 1997]. The genetic templates of the proteins are obtained from the ENA [Leinonen et al., 2011].

We compare the GC contents of the coding sequences of the proteins to the GC content of the corresponding genome, and to the genome the protein matching in the BLAST search originates from. As the virus host is an obvious source for horizontally transferred genes, also the genomes of the virus hosts are included in the comparison.
It is known that viruses exhibit a high mutation rate, and genes even from the same species might show a sequence similarity beyond levels of recognition. In order to obtain a reliable estimate of the GC content for an individual protein, as many nucleotide sequences as possible from the protein of interest are taken into account: In addition to the genes referenced in the UniProtKB entries from the complete proteomes, all genes referenced by UniProtKB entries that yield at least 95% similarity to the protein of interest are considered. Determination of the GC content on the basis of multiple sequences not only yields a more robust estimate, but also allows for assessment of statistical significance of the results. When comparing GC contents we conduct a Wilcoxon rank-sum test and provide the resulting p-value. The p-value states how likely the null hypothesis of the test holds. That is, how likely two sets of measurements are drawn from the same distribution.

**CodonW** CodonW is a tool for correspondence analysis of codon and amino acid usage. It also provides standard indices of codon usage, such as the GC3s and the GC content.

**European Nucleotide Archive** ENA collects and provides nucleotide sequencing data. Data records comprise nucleotide sequences as well as information such as experimental set up and machine configuration, output machine data, sequence traces, reads and quality scores, and interpreted information, such as assembly, mapping and functional annotation. Data from ENA is in constant exchange within the INSDC.

### 4.3 Results

Conducting a BLAST search for similar proteins from unrelated viruses, after pruning, we obtain 160 hits between viruses different in their nucleic acid type according to the Baltimore classification. The obtained proteins stem from all different types of viruses except dsDNA-RT. As depicted in figure 4.1, we find proteins from (+)ssRNA viruses to match with all other types of viruses except ssRNA-RT viruses. Proteins from dsDNA viruses match with proteins from ssRNA-RT and ssDNA viruses.

![Virus Types Matching](image)

Figure 4.1: Virus types of which proteins are found to match with each other in the BLAST result. Arrows indicate hits between proteins from the respective virus types.
### Table 4.1: Overview of the structural families observed in unrelated viruses. Viruses the matching proteins originate from are summarized in the last column. The second column denotes, if available, the PDB identifier of a representative of the respective family.

<table>
<thead>
<tr>
<th>Pfam family</th>
<th>Repr.</th>
<th># Hits</th>
<th>Matched virus types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parvo NS1</td>
<td>1U0J</td>
<td>1</td>
<td>Bocavirus (ssDNA) – Fowl adenovirus A (dsDNA)</td>
</tr>
<tr>
<td>MMTV SAg</td>
<td>-</td>
<td>2</td>
<td>Betaretrovirus (ssRNA-RT) – Rhadinovirus (dsDNA)</td>
</tr>
<tr>
<td>OrfB IS605</td>
<td>-</td>
<td>3</td>
<td>Inovirus (ssDNA) – Myoviridae, Bicaudavirus (dsDNA)</td>
</tr>
<tr>
<td>Pkinase Tyr</td>
<td>2IVS</td>
<td>3</td>
<td>Alpharetrovirus (ssRNA-RT) – Mimivirus (dsDNA)</td>
</tr>
<tr>
<td>RdRP I</td>
<td>2ECO</td>
<td>3</td>
<td>Crypovirus (dsRNA) – Potyviridae (ssRNA(+))</td>
</tr>
<tr>
<td>Hema esterase</td>
<td>3I27</td>
<td>4</td>
<td>Influenzavirus C ((-)ssRNA)–Coronaviridae ((+))ssRNA)</td>
</tr>
<tr>
<td>Helicase C</td>
<td>4C9B</td>
<td>5</td>
<td>Molluscipoxvirus (dsDNA) – Pestivirus ((+)ssRNA)</td>
</tr>
<tr>
<td>Phage integrase</td>
<td>1AIH</td>
<td>5</td>
<td>Inovirus (ssDNA) – Caudovirales (dsDNA)</td>
</tr>
<tr>
<td>RNA helicase</td>
<td>–</td>
<td>7</td>
<td>Norovirus ((+)ssRNA) – Circovirus (ssDNA)</td>
</tr>
<tr>
<td>Collagen</td>
<td>–</td>
<td>15</td>
<td>Sputnik (dsDNA) – Mimivirus, other dsDNA viruses *</td>
</tr>
<tr>
<td>dUTPase</td>
<td>1SYL</td>
<td>112</td>
<td>Betaretrovirus (ssRNA-RT) – several dsDNA viruses</td>
</tr>
</tbody>
</table>

*Our method reports this hit because the taxonomy entry of Sputnik virus is ‘unclassified virophage’ in the UniProt database.

The Pfam families of all pairs of proteins forming a hit are consistent and are summarized in table 4.1. This means, that for each pair of proteins in the BLAST result, the Pfam query yields the same protein family for both proteins. In total, the proteins are from 10 different Pfam families. The most prominent family is dUTPase. 70% of the proteins in the hit pairs belong to this family. dUTPase is a ubiquitous protein that hydrolyses deoxyuridine triphosphate (dUTP) to deoxyuridine monophosphate (dUMP) and pyrophosphate. The collagen family was identified only because of missing labels in the database: The *Sputnik virus* is recorded as ‘unclassified virophage’. But actually, *Sputnik virus* and *Mimivirus* both have a dsDNA genome. Moreover, *Mimivirus* acts as the host for *Sputnik virus*. Strong evidence for the HGT of genetic material between these two viral species is reported by La Scola and co-workers [La Scola et al., 2008], which supports the validity of our approach. We now discuss the findings for each Pfam family in detail.

#### 4.3.1 dUTPase Family

dUTPases are found in all domains of life. They catalyse the hydrolysis of dUTP to dUMP and pyrophosphate. Owing to their ubiquity, they are well studied and the 3D structures of multiple family members are known. The function of dUTPase is to regulate the cellular levels of dUTP in order to prevent incorporation of uracil in place of thymine into the DNA. dUTP is deoxygenated uridine triphosphate. Uridine, which is uracil bound to ribose, is a building block of RNA. By hydrolysis of dUTP to dUMP, dUTPase prevents its incorporation into the DNA. The hydrolysis product, dUMP, in turn can be processed to deoxythymine monophosphate (dTMP) which is needed to assemble DNA [Vértesy and Tóth, 2009].

dUTPase is a homotrimeric protein with each subunit consisting of a compact body and a tail. dUTP, the substrate of dUTPase, binds at the clefts between adjacent subunits with support of the tail from the opposing subunits (figure 4.2). Thus, all three subunits are involved in ligand
4.3 Results

Figure 4.2: Active site of *Mycobacterium tuberculosis* dUTPase. Substrate coordinating residues are labelled. Roman numerals stand for respective conserved motifs. Dashed lines indicate hydrogen bonds. Shaded rectangles indicate aromatic overlaps. In this dUTPase, the aromatic phenylalanine within motif 5 is replaced with a histidine [Vértesy and Tóth, 2009].

binding in each of the three binding sites. Each binding site is characterized by five motifs facilitating strict specificity for cleaving the $\alpha$-$\beta$ phosphate ester linkage of dUTPs. Their precise steric positioning precludes cleavage of deoxyuridine diphosphate (dUDP), and a specialized $\beta$-hairpin specifically associates with the uracil ring [Vértesy and Tóth, 2009]. Motif I, II, IV from one subunit (yellow in figure 4.2) contribute to binding the triphosphate chain. Motif III from an the adjacent subunit (blue in figure 4.2) primarily binds the nucleoside part of the dUTP molecule. Motif V from the tail of the opposite subunit (green in figure 4.2) has been suggested to conduct a nucleophilic attack on the R-phosphorus facilitated by closure of the arm upon the active site [Vértesy and Tóth, 2009].

We find dUTPase from *Betaretrovirus* to exhibit similarity to dUTPase of various dsDNA viruses. As mentioned above dUTPase is vital to all cellular organisms, and additionally it has been observed in dsDNA viruses and retroviruses [Barabás et al., 2006, Vértesy and Tóth, 2009, Chan et al., 2004]. Analysis of the alignments to the profile HMM of the dUTPase family and comparison to a resolved 3D structure of dUTPase from *Mycobacterium tuberculosis* indicate that the active site motifs of the protein are well conserved among all the obtained hits. An outline of the alignment containing the active sites is depicted in figure 4.3.

In the first motif, we observe aspartate in all aligned protein sequences. Aspartate is crucial to coordinate the magnesium ion [Vértesy and Tóth, 2009]. The second motif contains arginine to coordinate the ion and the phosphate chain. Arginine is also observed in all aligned sequences, except for one betaretroviral sequence (UniProt accession P21407) that shows little similarity in motif II. The third motif contributes primarily to binding the nucleoside. Tyrosine which coordinates the uracil ring and aspartartic acid which contributes a hydrogen bond we find in
Figure 4.3: Alignment of dUTPase proteins. Only the five conserved motifs are displayed. The first block are sequences from Betaretrovirus. The second block comprises sequences from dsDNA viruses. The last sequence is from Mycobacterium tuberculosis dUTPase of which the structure is resolved. Positions are provided corresponding to that sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.
4.3 Results

overall conserved. Only in some sequences tyrosine is replaced by phenylalanine which is also aromatic and thus can be assumed to preserve the coordinative function. Likewise, motifs IV and V are well conserved and we observe no crucial mutations in functionally important positions.

Despite the observed similarities, there is evidence that the function of dUTPase in retroviruses, in particular in *Mason-Pfizer monkey betaretrovirus* (UniProt accession P07570), is altered. Betaretroviral dUTPase is a bifunctional fusion protein that is found joint with a nucleocapsid protein and potentially contributes to folding of nucleic acid and to reverse transcription [Barabás et al., 2006]. The primary difference to other dUTPases is a shorter passage between the fourth and the fifth motif, yielding a shorter arm and hence low catalytic activity. Also, differences in the amino acid sequence compared to other dUTPases have been observed. Our analysis of the active site motifs however provides no evidence for altered functionality.

In mammalian herpesviruses, motif III occurs prior to motif I in the amino acid sequence. The altered arrangement is possibly caused by a gene duplication followed by point mutations. Remarkably, the active site architecture and the overall fold closely resemble the usual dUTPase trimeric fold [Vértessy and Tóth, 2009]. However, this rearrangement precludes our method from identifying herpesvirus dUTPase as a significant hit. This indicates that the method is tuned to identification of relatively close homologs.

4.3.2 RNA Helicase Family

RNA helicase unwinds RNA using the energy from adenosine triphosphate (ATP) hydrolysis. Five superfamilies of helicases have been identified, all of which bind ATP. ATP binding domains consist of characteristic motifs, called Walker A and Walker B [Walker et al., 1982]. The Walker A motif constitutes a phosphate binding loop. Lysine within this motif is essential for nucleotide binding. Coordination of the β- and γ-phosphates is achieved with the help of a magnesium ion coordinated by aspartic acid from the Walker B motif, while approximately 25

![Figure 4.4: Alignment of RNA helicase proteins in the positions of the characteristic motifs. The first three sequences are from *Circovirus*, the next three from *Norovirus*. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.](image)
residues upstream of the Walker A motif aromatic residues coordinate the adenine ring of ATP. Glutamate from Walker B is crucial for the hydrolysis of ATP [Snider and Houry, 2008].

We find the Replication-associated protein (Rep) of Circovirus (ssDNA) to exhibit similarity to P41 of Norovirus from the Caliciviridae family ((+)-ssRNA) and both proteins show characteristics of RNA helicase. A protein with resolved structure is not available for comparison. P41 is one of six proteins encoded in the polyprotein of Norovirus. P41 shares sequence motifs with protein 2C of Picornaviridae and superfamily 3 helicases. Experiments with purified P41 indicate nucleoside triphosphate (NTP) binding and NTP hydrolysis activity but not helicase activity [Pfister and Wimmer, 2001].

We observe a mutation in a crucial position of the Walker A motif for two of the three sequences from Circovirus, but well preserved motifs in the P41 sequences of Norovirus. As depicted in figure 4.4, at position 167 lysine, important for nucleotide binding, is replaced once by glutamine and once by arginine. While the change to glutamine is immense, the latter replacement theoretically might be function preserving as arginine and lysine have comparable properties.

The analysis of a further motif, characteristic for superfamily 3 helicases, confirms the similarity of P41 to this superfamily. Motif C consists of asparagine preceded by multiple hydrophobic residues. We find motif C in the proteins of Norovirus. But only one of the three sequences of Circovirus has the motif C defining asparagine residue in position 258. Thus, despite similarities to RNA helicases the function of the circoviral Rep might be altered.

Superfamily 3 helicases occur in small DNA and RNA viruses. Together with the origin binding domain, which is typically associated with the helicase in these viruses, superfamily 3 helicases allow viruses to replicate independently of regulation pathways of the host cell [Gorbalenya et al., 1989].

In conjunction with our observation of potential homology, Gibbs and Weiller hypothesize that a recombination event between Calicivirus and Circovirus endowed the Circovirus Rep with the helicase domains. Noticing similarities between the plant infecting Nanoviridae and the vertebrate infecting Circoviridae, the authors propose that Circoviridae evolved from Nanoviridae undergoing a host switch to vertebrates. As all known members of Caliciviridae, including Norovirus, infect vertebrates, the host switch presumably took place before the recombination event [Gibbs and Weiller, 1999].

### 4.3.3 Phage Integrase Family

Phage integrases are enzymes that mediate unidirectional site-specific recombination between two DNA recognition sequences, the phage attachment site (attP) and the bacterial attachment site (attB) [Groth and Calos, 2004]. After assembly of a nucleoprotein complex, the enzymatic reaction begins with the cleavage attB and attP. A specific tyrosine residue of phage integrase nucleophilically attacks the phosphate backbone. This tyrosine residue forms a covalent bond between the protein and the 3-phosphoryl group at the site of strand cleavage, leaving a free 5-hydroxyl group. Then, cleaved strands from adjacent DNA partners exchange positions with respect to each other, and the free 5-OH groups attack the opposite strands, releasing the bound integrase and relegating the cleaved strands [Hickman et al., 1997]. A crystal structure of C-terminal domain of the integrase protein of Bacteriophage HP1 (PDB accession 1AIH) which infects Haemophilus influenzae has been resolved [Hickman et al., 1997] and serves as a structural representative for our analysis.
4.3 Results

Figure 4.5: Excerpt of the structure of *Bacteriophage HP1* integrase (PDB accession 1AIH) with the catalytic sites highlighted.

Figure 4.6: Alignment of phage integrase proteins. The first sequence is from *Pseudomonas phage Pf1*. The sequences in the next block stem from members of Caudovirales. The last sequence is from the integrase protein of *Bacteriophage HP1*, the structure of which is known. Positions are provided corresponding to that sequence. Strictly conserved sites with literature support are marked by an arrow. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.
Proteins we find by the BLAST search with the characteristics of phage integrase are from *Pseudomonas phage Pf1* (ssDNA) and from members of Caudovirales (dsDNA) which are also phages. The alignment to the Pfam family profile, depicted in figure 4.6, shows that the tyrosine residue, which is crucial for the DNA cleavage, is conserved throughout all sequences. As visualized in the ribbon diagram in figure 4.5, three further strictly conserved basic residues are found in all integrases in proximity to the tyrosine residue. Possible roles of these two arginine and one histidine residues are orienting the DNA in the appropriate conformation for the nucleophilic attack, stabilizing a pentacoordinate transition state at the scissile phosphate, or participating in shuttling protons [Hickman et al., 1997]. Two of the three positions we find to be conserved in all obtained proteins. One, arginine at position 40 of the structural representative, we only observe in four out of the six sequences. The other two sequences from the Caudovirales species *Haemophilus phage HP1* and *Enterobacteria phage WPhi* show asparagine and alanine in the respective position.

Since all the organisms in which we observed the phage integrases are bacteriophages, and bacteria are known to share genes extensively, we assume that the observed similarity of these proteins is due to their common ancestry. And at least in four of six cases the proteins are likely to be fully functional.

### 4.3.4 Helicase C Family

The Helicase C family describes the C-terminal domain found in proteins belonging to helicase superfamily 2. The ATP binding site which is common to helicases, and has the characteristic motifs Walker A and Walker B, is not part of this Pfam family. Three motifs define the C-terminal region of superfamily 2 helicase, motifs IV, V and VI. Motif IV however is not included in the seed alignment provided by Pfam. The most prominent characteristic of superfamily 2 helicases is a strictly conserved arginine residue in the center of motif VI. This residue is possibly involved in RNA binding and is not found in the otherwise similar superfamily 1 helicases [Caruthers et al., 2000].

We find similar proteins from *Molluscipoxvirus* (dsDNA) and *Pestivirus* ((+)+ssRNA) that belong to this Pfam family. The alignment to the profile derived from the seed alignment for helicase C is shown in figure 4.7. We observe motif VI with the strictly conserved arginine in the center in the sequences of both viruses. In motif V, the first an the fourth position deviate from the seed alignment in both viral sequences, the second and third position of motif V are well conserved in the amino acid sequence of *Pestivirus*. In the sequence of *Molluscipoxvirus*, we observe serine in the second position of motif V while glycine is most common in this position. Glycine and serine are both small and polar. Thus, the protein sequences from both viruses indicate consistent chemical properties at this site. Motif IV of the structural representative we exclude from further analysis as the seed alignment only begins after this motif (at position 295 in figure 4.7).

From the analysis of the sequence alignment, we conclude that both viral sequences show similarity to the C terminus of helicase superfamily 2. The proteins might be functionally active. Helicases are indispensable for replication, transcription, and replication. They occur in all cellular organisms and hence unsurprisingly in many viruses.
4.3 Results

Figure 4.7: Alignment of the C-terminal domain of helicase superfamily 2 proteins. The first sequence is from Pestivirus, the next sequence from Molluscipoxvirus, and the last one, of which the 3D structure is known, is from human. Four sequences from Pestivirus, nearly identical to the one shown, are omitted. Positions and annotations of motifs are provided with respect to the last sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.
4 Detection of Horizontal Gene Transfer Between Viruses

4.3.5 Hema Esterase Family

The Hema Esterase Pfam family represents the core of the haemagglutinin-esterase glycoprotein, including the haemagglutinin receptor-binding domain as well as the esterase domain. The haemagglutinin-esterase glycoprotein is a trimer, where each monomer is composed of three domains: an elongated stem active in membrane fusion, an esterase domain, and a receptor-binding domain [de Groot, 2006].

The HEF is embedded in the viral envelope. Originally discovered in Influenzavirus C, it was also found in Coronaviridae, namely in Torovirus and in group 2 coronaviruses [de Groot, 2006]. It recognises the host cell surface receptor, fuses the viral and the host cell membranes, and destroys the receptor upon host cell infection.

The hits we find match the virus genera for which the presence of HEF has been described before: Influenzavirus C ((-)ssRNA), and Torovirus and Betacoronavirus both from Coronaviridae ((+)-ssRNA). The alignment, depicted partially in figure 4.8, shows well conserved active sites for both the receptor binding site and the enzyme active site. An illustration of the enzyme active site is shown in figure 4.9. The site is composed of serine (position 57), histidine (355), and aspartate (352). The NH groups of glycine (85) and serine (57) together with the side chain of asparagine (117) form the oxyanion whole. Arginine (322) forms hydrogen bonds with the substrate [de Groot, 2006].

The source of the HEF gene is unknown. The HEF proteins of toroviruses, coronaviruses and influenza C viruses are evolutionary equidistant with about 30% sequence identity in all directions. The gene might have originated in any of the three viruses and subsequently have spread to the others via horizontal gene transfer [de Groot, 2006].

![Figure 4.8: Alignment of haemagglutinin esterase proteins. The first block of sequences is from Coronaviridae. The last sequence of this block stems from a partial protein. The next sequence is from Influenzavirus C, and the last sequence, of which the structure is known, is from Torovirus. Positions and annotations of motifs are provided with respect to the last sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.](image-url)
Figure 4.9: Receptor binding site of HEF with ligand [Rosenthal et al., 1998]. The catalytic residues are shown in green.

4.3.6 RdRP 1 Family

RNA directed RNA polymerase is an essential protein encoded in all genomes of RNA containing viruses with no DNA stage [Koonin et al., 1989]. It catalyses synthesis of the RNA strand complementary to a given RNA template, but the precise molecular mechanism remains unclear. The postulated RNA replication process is a two step mechanism involving a primer independent initiation and an elongation phase [Kao et al., 2001]. All RNA containing viruses with no DNA stage belong to four major classes corresponding to the type of polymerases they encode [O’Reilly and Kao, 1998].

The structure of RNA directed RNA polymerase resembles that of other polymerases: a right hand with fingers, palm, and thumb subdomains (PDB accession 2ECO). In addition, there is an N-terminal segment that encircles the active site of the enzyme bridging the fingers and the thumb subdomain. This region, termed the fingertips, is unique to RNA directed RNA polymerases compared to other polymerases [Ferrer-Orta et al., 2004]. A ribbon diagram of the protein is depicted in figure 4.11.

The palm subdomain which is a conserved feature of all polymerases contains five conserved motifs, A to E. The motifs are involved in nucleotide binding, phosphoryl transfer, structural integrity of the palm subdomain, and priming nucleotide binding. Studying this structure, Ferrer-Orta and co-workers find 27 amino acids to be in direct contact with RNA residues. Eleven of them are from the conserved structural motifs B, C, and E in the palm and F in the fingers, and the remaining 16 residues belong to different regions on the N terminus, fingers and thumb subdomains [Ferrer-Orta et al., 2004].

The Pfam family RdRP 1 only comprises the motifs A and B from the palm subdomain and the fingers towards the N terminus where motif F is located. In figure 4.10 the alignment in the positions of these motifs is depicted. In the BLAST search result we observe hits between Cryspovirus (dsRNA) and three members of Potyviridae ((+))ssRNA) showing characteristics
Detection of Horizontal Gene Transfer Between Viruses

Figure 4.10: Alignment of RNA dependent RNA polymerase proteins. The first three sequences are from members of Potyviridae, the next from Cryspovirus and the last from Foot-and-mouth disease virus of which the structure is resolved. Positions and annotations of motifs are provided with respect to the last sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.

of this Pfam family. The potyviral sequences coincide well with the conserved motifs. For the sequence from Cryspovirus (UniProt accession O15925) however, motif F appears to be altered leaving room for speculation whether the functionality is preserved and whether this protein belongs to this subgroup of polymerases.

Judging from the virus types, all found proteins are from the first of the four classes of RNA containing viruses with no DNA stage, comprising viruses containing (+)ssRNA or dsRNA, except retroviruses and Birnaviridae. Although RNA dependent RNA polymerases are present in the genome of all RNA containing viruses with no DNA stage, we observe only three hits. This fact possibly attributes to the low overall sequence homology among RNA dependent RNA polymerases. The domain organization and the 3D structure of the catalytic centre however are conserved [O’Reilly and Kao, 1998].
4.3 Results

Figure 4.11: Structure of RNA dependent RNA polymerase bound to RNA oligonucleotide (PDB accession 2ECO). Fingers, palm, and thumb refer to the names of the subdomains. The motifs F (magenta), A (red), and B (blue) are highlighted.

4.3.7 Pkinase Tyr Family

Protein kinases mediate protein phosphorylation, one of the most important posttranslational modifications of proteins. They catalyse the transfer of the phosphate from nucleotide triphosphates, such as ATP, to amino acid residues in a protein substrate side chain. The result is a conformational change affecting protein function. Protein kinases fall into three broad classes, characterised with respect to substrate specificity [Hanks et al., 1988]: Serine/threonine-protein kinases, tyrosine protein kinases, and dual specificity protein kinases.

The Pkinase Tyr family represents the catalytic domain found in a number of serine/threonine and tyrosine protein kinases. It does not include catalytic domain of dual specificity kinases. We observe hits between proteins from Mimivirus (dsDNA) and Alpharetrovirus showing characteristics of this family. The protein sequences in particular match with the conserved regions of the seed alignment as illustrated in figure 4.12. The representative with resolved structure, included for comparison, is a human protein to which the viral proteins, in particular the retroviral protein, show similarity.

In conjunction with our hypothesis, that the found viral Pkinase Tyr proteins are homologous and exhibit similar function, it has been shown that both these viruses encode protein kinases [Raoult et al., 2007, Jacob et al., 2011]. Jacob and co-workers however point out that virus-encoded serine/threonine protein kinases appear to be a feature that is unique to large DNA viruses. Oncogenic tyrosine kinases encoded by acute transforming retroviruses are well characterized viral tyrosine kinases. However, these kinases are captured forms of cellular genes, and their incorporation may actually be considered an accident during virus replication, typically indicating defective virus replication [Jacob et al., 2011]. Thus the similarities that we observe between viral proteins are probably caused by a deeper relatedness between all cellular serine/threonine and tyrosine kinases.
Figure 4.12: Alignment of serine/threonine-protein kinase. The first sequence is from Alpharetrovirus, the next three sequences from Mimivirus and the last from human of which the structure is resolved. Positions are provided with respect to the last sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.
4.3 Results

4.3.8 OrfB IS605 Family

OrfB IS605 is a putative domain found at the C-terminus of a large number of transposase proteins. Transposases are enzymes that bind to the ends of transposons, DNA sequences that can change its position within the genome, and catalyse the movement of the transposon to another part of the genome. OrfB IS605 is found in a number of uncharacterised bacterial proteins. For instance, an insertion sequence-like element (IS1341) of the Thermophilic bacterium PS3 belongs to this Pfam family exhibiting high similarity to IS891 from the cyanobacterial species Anabaena, and to IS1136 from Saccharopolyspora erythraea [Murai et al., 1995].

We observe this Pfam family for a hit between Enterobacteria phage IfI (ssDNA) and Clostridium botulinum C phage (dsDNA), as well as for a hit between Enterobacteria phage IfI (ssDNA) and Acidianus two-tailed virus (dsDNA). The latter virus infects Acidianus convivator, an archaean species.

Little appears to be known about this Pfam family and there is no resolved structure available to compare with. Analysis of the alignment to the profile shows that the majority of motifs are preserved in the viral proteins, suggesting that they also encode transposase.

The profile derived from the seed alignment of this family shows various regions of high conservation. We align the found proteins to the profile and analyse the alignment focussing on the regions of high conservation: Position 24, 27, and 28 show a preference aromatic amino acids such as tyrosine, tryptophane, or phenylalanine, sometimes substituted by non-polar amino acid

Figure 4.13: Alignment of OrfB IS605 proteins. The first sequence is from Enterobacteria phage IfI, the next two sequences are from Clostridium botulinum C phage, and the last sequence is from Acidianus two-tailed virus. Positions are provided with respect to the last sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.
such as leucine. At position 49, all sequences in the profile have proline. Position 54 is again aromatic. Likewise position 150 shows a preference for aromatic amino acids, followed by an accumulation of isoleucine, valine, and cysteine in the subsequent positions, 152, 154, and 156. A highly conserved pattern is found at position 192-198. Aspartic acid forms the center of this motif and is surrounded by non-polar amino acids like leucine, methionine, and isoleucine. These three amino acids are surrounded by glycine or other polar amino acids on either side, and again by non-polar amino acids. Position 260 and 263 show a preference for histidine, aspartic acid, or asparagine. Another accumulation of non-polar amino acids is found at positions 269, 270, 278, 281, 283, and 286.

As shown in figure 4.13, we find overall good resemblance of the described pattern in the aligned viral sequences. In particular, the very conserved pattern with aspartic acid at position 195 is present in all sequences. The proline at position 49 however is only present in two of the four viral sequences. One of the sequences of *Clostridium botulinum C phage* and the protein of *Acidianus two-tailed virus* do not show proline at position 49 but isoleucine and leucine. Furthermore, the sequence of *Acidianus two-tailed virus* has arginine at position 152 instead of a non-polar amino acid. And likewise, the sequence of *Enterobacteria phage IfI* has serine at position 270 instead of a non-polar amino acid. Due to the observed similarities, and the presumed intensive sharing of genetic material between ancestors of bacteria and their phages, we hypothesize a common origin for the described proteins.

Transposases use many different mechanisms for cutting and pasting DNA. Other transposases have been described in bacteria (e.g. PDB accession 1F3I) [Davies, 2000] and bacteriophages (e.g. 1Z1G) [Biswas et al., 2005]. Similarly, retroviral integrase as of HIV catalyses a transposition event [Craig, 1995]. None of these appears to be related to the OrfB IS605 family.

4.3.9 MMTV SAg Family

Superantigens are proteins of microbial origin and cause non-specific activation of T cells resulting in activation of polyclonal T cells and massive release of cytokines. They are a defence mechanism against the adaptive immune system triggering an immense immune response undermining the ability of the adaptive immune system to target antigens with high specificity [Acha-Orbea and MacDonald, 1995].

The MMTV SAg Pfam family describes a superantigen of *Mouse mammary tumor virus*, a milk-transmitted betaretrovirus. A structure of this protein is not available. Also, the seed alignment does not allow for comparison as it is based on only five sequences which are very similar to one another. Thus it is not possible to distinguish between regions of high and low conservation.

The hits we observe are between proteins from *Mouse mammary tumor virus* (ssRNA-RT) and *Herpesvirus ateles* as well as *Herpesvirus saimiri* both from the genus *Rhadinovirus* (dsDNA). While the sequence from *Mouse mammary tumor virus* well resembles the sequences in the seed alignment, the proteins from *Rhadinovirus* cannot be compared in a meaningful way. The *Herpesvirus ateles* protein is labelled 'Mitogen' indicating that it might play a role in triggering mitosis and hence cell proliferation. The *Herpesvirus saimiri* has been subject of a study indicating that T cell proliferation upon infection is ascribed to that gene [Knappe et al., 1997]. Thus, at least the in the latter case the observed similarity may indicate a functionally homologous and potentially evolutionary related protein.
4.3 Results

4.3.10 Parvo NS1 Family

Parvo NS1 represents the helicase domain of the parvoviral NS1 protein. This protein is required for DNA replication. It has ATPase activity and also regulates host gene expression through histone acetylation [Iseki et al., 2005].

Interestingly, this Pfam family is found not only in *Bovine parvovirus* (ssDNA), but also in *Aviadenovirus* (dsDNA). Comparison of the two proteins found by BLAST to the seed alignment shows good resemblance. However, the protein from *Aviadenovirus* is part of the seed alignment, posing the question how the sequences for this alignment are selected. There appears to be no literature support that the Pfam family Parvo NS1 specific to *Parvoviridae* has been studied outside this viral family. However, considering the alignment without sequences of *Aviadenovirus* yields a consistent picture. The conservation of the active sites is evaluated using information from a structural representative (PDB accession 1U0J). It is derived from *Adeno-associated virus* 2, a parvovirus that can replicate only under co-infection of other viruses, such as adenovirus [Carter et al., 1973]. The ATP binding site of NS1 of the representative structure is formed by Walker A and B motifs at positions 111-119 and 231-242. These sites are depicted in figure 4.14. The structure, although showing characteristics of the Parvo NS1 family, has been identified as a superfamily 3 helicase [James et al., 2004].

The first part of this site is well conserved among the sequences in the seed alignment. The *Bovine parvovirus* protein from the identified pair is almost identical to the structural representative in this part, except one replacement of threonine by serine. The *Aviadenovirus* protein matches only two glycine residues which are primarily defining the Walker A motif that facilitates ATP binding. The lysine after the second glycine (position 117) followed by serine or threonine is also characteristic for Walker A, but is not found in the adenoviral sequence. The Walker B motif shows more variability among the sequences in the seed alignment from parvoviruses, which prevents from drawing conclusions how well the adenoviral protein matches the pattern.

Figure 4.14: Alignment of Parvo NS1 proteins. The first sequence is from *Aviadenovirus*, the second from *Bovine parvovirus*, and the last, of which the structure is known, from *Adeno-associated virus* 2, a member of the *Parvoviridae* family. Positions an motif annotations are provided with respect to the last sequence. The shading discriminates between acidic (red) and basic (blue), polar uncharged (yellow) and hydrophobic nonpolar (green) residues.
Detection of Horizontal Gene Transfer Between Viruses

Provided the fact that Bovine parvovirus can only replicate under co-infection with adenoviruses (or other viruses, such as Herpes simplex virus [Handa and Carter, 1979]), a homology of the helicases appears likely. Helicases are known to exhibit a high level of sequence variability [Gorbalenya and Koonin, 1993], thus the adenoviral protein which is not characterized further in the UniProt database might exhibit helicase activity.

4.3.11 GC Content Analysis

To find evidence whether HGT between the viruses that yield hits in the BLAST search is a possible explanation for their similarity, the GC contents are compared. We analyse the GC content of the genetic code of each protein, of the corresponding complete genome, and of the corresponding virus host. As synonymous codons primarily differ in the third codon position, we use only this position for the GC content calculation, which is referred to as GC3s.

According to statistical analysis, a number of genes appear to better fit the genome of the matching protein than the own genome. We observe the genes of many proteins to be significantly dissimilar from their corresponding genome and similar to the genome of the matching protein using a significance level of 0.05 tested with Wilcoxon ranksum test. These proteins are displayed in table 4.3. However, as this table shows, the absolute differences in GC3s content are small, such that this analysis provides little evidence whether we see examples of HGT.

The genetic sequences of some proteins exhibit comparably large deviations from the virus GC3s content, but are not necessarily closer to the matching proteins genome GC3s content. These proteins are listed in table 4.2 and could be suspected to have a foreign source we did not analyse.

One obvious source of horizontally transferred genes for viruses is their host. Viral genes with GC3s content noticeably similar to that of the host are listed in table 4.4. In favour of the hypothesis that some viral genes originate from the host, a sequence alignment for dUTPase (UniProt accession Q6TVW3) from Orf virus to human dUTPase yields 72% sequence identity, and an alignment between ORF348 (O80301) from Enterobacteria phage If1 to IS605 OrfB family transposase from Escherichia coli yields 99% identity. The horizontal transfer of the gene for dUTPase from the host to the virus has been subject of a previous study [Baldo and McClure, 1999].

Table 4.2: Proteins with largest deviations in GC3s content from the corresponding genome.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Family</th>
<th>Protein GC3s</th>
<th>Virus GC3s</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00526</td>
<td>Pkinase Tyr</td>
<td>0.795</td>
<td>0.596</td>
<td>0.0000</td>
</tr>
<tr>
<td>O80301</td>
<td>OrfB IS605</td>
<td>0.507</td>
<td>0.360</td>
<td>-</td>
</tr>
<tr>
<td>Q6TVW3</td>
<td>dUTPase</td>
<td>0.522</td>
<td>0.854</td>
<td>0.0019</td>
</tr>
<tr>
<td>P21407</td>
<td>dUTPase</td>
<td>0.494</td>
<td>0.339</td>
<td>-</td>
</tr>
<tr>
<td>Q9T1R0</td>
<td>Phage integrase</td>
<td>0.250</td>
<td>0.373</td>
<td>0.0018</td>
</tr>
<tr>
<td>O41138</td>
<td>Collagen</td>
<td>0.254</td>
<td>0.397</td>
<td>-</td>
</tr>
<tr>
<td>Q857L6</td>
<td>Collagen</td>
<td>0.735</td>
<td>0.845</td>
<td>0.0023</td>
</tr>
<tr>
<td>Q8VAA1</td>
<td>Collagen</td>
<td>0.226</td>
<td>0.376</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Table 4.3: Proteins significantly dissimilar in GC3s content from their genome and similar to the genome of the matching protein. The difference between the GC3s content of the protein and its corresponding genome is given in column 'Diff own'. The difference between the GC3s content of the protein and the genome of the matching protein is given in column 'Diff other'. The p-values are obtained by Wilcoxon rank-sum test tests, testing the null hypothesis that two sets of measurements are drawn from the same distribution. The sets of measurements are as described in 4.2.4.

<table>
<thead>
<tr>
<th>Pfam family</th>
<th>Query and subject protein</th>
<th>Diff own</th>
<th>p-Val</th>
<th>Diff other</th>
<th>p-Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collagen</td>
<td>Q5UQ50 Mimivirus (dsDNA)</td>
<td>0.073</td>
<td>0.000</td>
<td>0.007</td>
<td>0.770</td>
</tr>
<tr>
<td></td>
<td>B4YNE6 unclassified virophages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collagen</td>
<td>Q5UPS7 Mimivirus (dsDNA)</td>
<td>0.055</td>
<td>0.000</td>
<td>0.011</td>
<td>0.672</td>
</tr>
<tr>
<td></td>
<td>B4YNE6 unclassified virophages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA helicase</td>
<td>Q83883 Norovirus ((+)-ssRNA)</td>
<td>0.051</td>
<td>0.000</td>
<td>0.011</td>
<td>0.559</td>
</tr>
<tr>
<td></td>
<td>Q805H4 Circovirus (ssDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA helicase</td>
<td>Q04544 Norovirus ((+)-ssRNA)</td>
<td>0.051</td>
<td>0.000</td>
<td>0.011</td>
<td>0.559</td>
</tr>
<tr>
<td></td>
<td>Q805H4 Circovirus (ssDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA helicase</td>
<td>P54634 Norovirus ((+)-ssRNA)</td>
<td>0.051</td>
<td>0.000</td>
<td>0.011</td>
<td>0.559</td>
</tr>
<tr>
<td></td>
<td>Q805H4 Circovirus (ssDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA helicase</td>
<td>Q83883 Norovirus ((+)-ssRNA)</td>
<td>0.051</td>
<td>0.000</td>
<td>0.016</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>Q8BB16 Circovirus (ssDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA helicase</td>
<td>Q04544 Norovirus ((+)-ssRNA)</td>
<td>0.051</td>
<td>0.000</td>
<td>0.016</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>Q8BB16 Circovirus (ssDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collagen</td>
<td>Q5UPX3 Mimivirus (dsDNA)</td>
<td>0.047</td>
<td>0.001</td>
<td>0.019</td>
<td>0.313</td>
</tr>
<tr>
<td></td>
<td>B4YNE6 unclassified virophages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P51518 Betaretrovirus</td>
<td>0.037</td>
<td>0.001</td>
<td>0.009</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Q48500 Bacteriophage T5 (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P51518 Betaretrovirus</td>
<td>0.037</td>
<td>0.001</td>
<td>0.009</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Q66LT4 Bacteriophage T5 (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P51518 Betaretrovirus</td>
<td>0.037</td>
<td>0.001</td>
<td>0.011</td>
<td>0.487</td>
</tr>
<tr>
<td></td>
<td>Q4A282 Coccolithovirus (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>Q4A282 Coccolithovirus (dsDNA)</td>
<td>0.037</td>
<td>0.020</td>
<td>0.011</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>P51518 Betaretrovirus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P07570 Betaretrovirus</td>
<td>0.034</td>
<td>0.023</td>
<td>0.009</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Q48500 Bacteriophage T5 (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P04024 Betaretrovirus</td>
<td>0.034</td>
<td>0.023</td>
<td>0.009</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Q48500 Bacteriophage T5 (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P07570 Betaretrovirus</td>
<td>0.034</td>
<td>0.023</td>
<td>0.009</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Q66LT4 Bacteriophage T5 (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dUTPase</td>
<td>P04024 Betaretrovirus</td>
<td>0.034</td>
<td>0.023</td>
<td>0.009</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>Q66LT4 Bacteriophage T5 (dsDNA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4 Detection of Horizontal Gene Transfer Between Viruses

Table 4.4: Proteins with GC3s content at most 0.05 different from that of the virus host and with a GC3s content not significantly similar to the virus genome GC3s content.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Family</th>
<th>Protein GC3s</th>
<th>Host GC3s</th>
<th>Virus GC3s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q38067</td>
<td>Phage integrase</td>
<td>0.742</td>
<td>0.786</td>
<td>0.721</td>
</tr>
<tr>
<td>Q33211</td>
<td>OrfB IS605</td>
<td>0.124</td>
<td>0.133</td>
<td>0.143</td>
</tr>
<tr>
<td>Q331V1</td>
<td>OrfB IS605</td>
<td>0.124</td>
<td>0.133</td>
<td>0.143</td>
</tr>
<tr>
<td>O80301</td>
<td>OrfB IS605</td>
<td>0.507</td>
<td>0.532</td>
<td>0.360</td>
</tr>
<tr>
<td>Q38106</td>
<td>dUTPase</td>
<td>0.259</td>
<td>0.248</td>
<td>0.226</td>
</tr>
<tr>
<td>Q6TVW3</td>
<td>dUTPase</td>
<td>0.552</td>
<td>0.554</td>
<td>0.854</td>
</tr>
<tr>
<td>Q857L6</td>
<td>Collagen</td>
<td>0.735</td>
<td>0.731</td>
<td>0.845</td>
</tr>
<tr>
<td>Q9XJK8</td>
<td>Collagen</td>
<td>0.502</td>
<td>0.551</td>
<td>0.493</td>
</tr>
</tbody>
</table>

4.4 Discussion

Proteins from unrelated viruses which exhibit similarity were collected by means of a BLAST search. The obtained proteins were of ten different protein families, each of which we studied intensely. Alignments to profiles of each family and comparison to proteins with resolved structure of the respective family indicate homology between the pairs of similar proteins. For many proteins we hypothesise not only a common evolutionary origin but also preserved functionality. A GC content analysis of the obtained proteins does not provide clear evidence for HGT between the viral species. However, we observe clear indications of transfer between host and virus.

The proteins we find to exhibit similarities across different virus types act predominantly on the processing or modification of nucleic acid. The ability to replicate on the basis of nucleic acid is the unifying characteristic of all organisms. For replication, all need to process this material, or in case of viruses, need to convince their host to process for them: Helicases, proteins that unpack genes, are ubiquitous to all cellular organisms. We find different types of them among viruses. For instance, RNA based viruses need to provide an own version of helicase to the host cell, suitable for their genome. dUTPases, also vital to all cellular organisms, prevent the incorporation of uracil in the DNA. Transposases catalyse the movement of transposable elements within a genome. Phage integrases, not found in cellular organisms, mediate the recombination between phage DNA and bacterial DNA.

Of the remaining Pfam families, only Pkinase Tyr has a function important to all organisms. The other two, MMTV SAg and in particular Hema Esterase, are very specific to certain viruses. For both it would be very interesting to investigate further on their origins. How is it possible that these same Pfam families occur in unrelated viruses? Gene transfer from a (maybe past) common host is not a likely possibility as there is no cellular organism with these genes. Options how these genes could have been attained are by a mechanism of gene exchange between viruses, by common ancestry or a recombination event, or by parallel evolution.

The initial question, whether there is exchange of genetic material between viruses of different types, cannot be affirmed. The analysis of GC contents provides evidence for HGT from host to virus for some genes considered in this study, but no clear evidence for gene transfer between viruses. Most proteins found to exhibit similarities across different virus types could have
well been obtained from hosts. Many of these proteins are ubiquitous in all cellular organisms, thus not even a common host is necessary to provide viruses with similar proteins.

Events of HGT in viruses for genes of the replication machinery have been reported previously. Moreira finds evidence for multiple horizontal transfers of these informational genes from bacteria to their specific phages and assumes it to be a common phenomenon [Moreira, 2000]. This transfer, he points out, implies an acceleration of evolutionary rates of the transferred genes, because phages normally evolve faster than their cellular hosts.

Rates of evolution might also be the reason, why we observe similarities primarily between informational genes of unrelated viruses. We did not observe any proteins forming the virus scaffold, such as the capsid or the nucleocapsid, to be similar between unrelated viruses. These proteins rarely have relatives outside their immediate family, which might be attributed to a more rapid evolution of this type of proteins, making relatedness undetectable on sequence level. A future direction would be to incorporate structural features in the analysis, opening the possibility to find more distant relatives.
5 Model to Detect Horizontal Gene Transfer Into Viruses

In this chapter, we introduce and evaluate a method to identify genes atypical for the virus family they were observed in. The central element of the method is a one class SVM which, given a statistical signature of a gene, predicts whether the gene is atypical for the family. We evaluate the method for a variety of different input features derived from the genetic sequences and compare their performance. The evaluation of the method is two-fold: We assess prediction success on simulated data and on real data. In the latter case, the enrichment of foreign genes in BLAST searches for the prediction results serves as ground truth data to compare to. Finally, for selected virus families we consult literature and discuss whether the found atypical genes are examples of horizontally transferred genes.

5.1 Problem Setting

Genetic sequences of a genome are assumed to exhibit common statistical features specific to that genome (compare section 2.6), whereas genes (recently) horizontally transferred into the genome show the specific statistical features of their original genome. Given all genes of a viral family, we use genomic signatures and aim to identify those genes that do not originate from any species in that family. The signature must be constructed such that it distinguishes typical from atypical genes. And ideally it is so simple and intuitive that we could tell from looking at it whether a gene is foreign to a family. This however is beyond all expectations. What we can hope for is to find a description with which a learning method can solve the task.

Still this is a hard problem. Representations of genes in the form of signatures that allow for distinction of within-family and out-of-family genes are based on statistical measures of the nucleotide composition of genes. But reliable statistics of viral genes are hard to obtain. Fast evolutionary rates and short genomes make statistical assessment very difficult. Hence, representations might be affected by high fluctuations and potentially exhibit large statistical errors. In particular for high dimensional representations, biases might be extremely high, completely hiding the actual information. Options to overcome this are either to work with simple representations or to find a way to reduce biases. Representations obtained from either option are evaluated as inputs for a learning method. We have used a trick to reduce biases adapted from metagenomics studies.

**Inspiration from Metagenomics Research** Scientists in metagenomics have to meet similar challenges. Their task is to extract and identify genetic material from environmental samples. Such samples are a heterogeneous mixture of short fragments of DNA from various species and the challenge is either to map them to the matching species or to identify them as novel. This identification requires statistical analysis of the nucleotide composition. Statistical assessment however is more stable and reliable if the data base is large. Particularly, codon frequencies observed in short sequences might exhibit large biases. To overcome this, the idea implemented by
5 Model to Detect Horizontal Gene Transfer Into Viruses

Zhu et al. [Zhu et al., 2010] is to use dependencies apparently formed in evolution: A linear relation exists between the codon frequencies and the genome nucleotide composition. Therefore, an estimate of the true codon frequencies can be reconstructed from frequencies of nucleotides observed in a short sequence. This idea is re-utilized in the approach described subsequently.

5.2 Approach

To predict if genes potentially have alien origin, a model is build. The construction of the model involves two major parts: First, the preparation of input features, and second, the training of the model using these features. The employed model is a one class SVM, an unsupervised learning algorithm which is discussed in detail in this section. The last part of this section describes the methodology used for evaluation of the constructed model. This evaluation is particularly demanding because evaluation procedures for learning algorithms typically involve knowledge about ground truth data, but in the case of horizontal gene transfer, these data are available only to a very limited extend.

5.2.1 Prediction Algorithm

The identification of potentially horizontally transferred genes resembles an outlier detection problem. We hypothesize that most genes of a family are inherent and only few entered the family by means of horizontal transfer. Furthermore, the family-inherent genes share characteristics which the foreign genes do not have. A one class SVM is particularly suited for this task. It is an unsupervised algorithm that learns a decision function for outlier detection. New data provided to the algorithm is classified as similar or different to the before seen data.

Supervised and Unsupervised Learning

Machine learning algorithms are methods that, from examples of data presented, generalize to a model which subsequently allows to evaluate new data instances. Two major types of machine learning methods are distinguished: supervised and unsupervised learning. In supervised learning an algorithm is provided with input data $X$, where for each sample $x_i$ an expected output $y_i$ is known during the training phase. The set of pairs of input and expected output is called training set. With the help of this set, a supervised learning algorithm can adjust parameters such that the actual outputs $\hat{Y}$ are similar to the expected outputs $Y$. For new samples $X'$ that were not seen during training, it then can predict an the output $\hat{Y}'$.

In unsupervised learning, in contrast, there is no expected output $Y$ to train the algorithm. The training data is called unlabelled. As there is no true output to compare the result of the algorithm to, this also implies that the success of the prediction cannot be measured directly. For this problem, Hastie et al. [Hastie and Tibshirani, 2009] remark that the assessment of the quality of the results typically needs to be based on heuristic arguments and the effectiveness of such algorithms is rather a matter of opinion than of rigorous verification. Nevertheless, this class of algorithms is inevitable when labelled data is unavailable.
Support Vector Machines

Support vector machines [Boser et al., 1992] are typically employed as a supervised learning approach. To understand how they can also be used in the unsupervised setting, it is helpful to first discuss the supervised case for binary classification.

Given features \(x_1, x_2, \ldots, x_N\) from the feature space \(\mathbb{R}^P\) and their labels \(y_1, y_2, \ldots, y_N \in \{-1, +1\}\), the task of an SVM is to find the optimal hyperplane separating the features with label \(+1\) from the features with label \(-1\). The best separating hyperplane is the one that has maximal distance to the closest features from both classes, a maximum margin classifier.

With the hyperplane represented by \(\{x : f(x) = \omega x + \omega_0 = 0\}\), where \(\omega \in \mathbb{R}^N\) is a unit vector with \(||\omega|| = 1\) and \(\omega_0 \in \mathbb{R}\), an optimization problem can be formulated to obtain the hyperplane with maximum margin and the two classes perfectly separated:

\[
\min_{\omega, \omega_0} \frac{1}{2}||\omega||^2 \\
\text{subject to } y_i(x^T_i \omega + \omega_0) \geq 1, \quad i = 1, \ldots, N
\]

As it is not necessarily possible to find a perfect separation between the two classes of features, slack variables \(\xi = (\xi_1, \xi_2, \ldots, \xi_N)\) are introduced to allow for some samples to be on the wrong side of the margin. They allow but penalize false classifications in the optimization. With the cost parameter \(C\) trading off between margin maximization and minimization of the number of false classifications, the optimization problem becomes

\[
\min_{\omega, \omega_0} \frac{1}{2}||\omega||^2 + C \sum_{i=1}^{N} \xi_i \\
\text{subject to } y_i(x^T_i \omega + \omega_0) \geq 1 - \xi_i \\
\xi_i \geq 0, \quad i = 1, \ldots, N
\]

and is often referred to as soft margin SVM.

A dual form of this optimization problem can be derived by reformulation using Lagrangian multipliers [Vapnik, 1995]. The Lagrange primal form is

\[
L_P = \frac{1}{2}||\omega||^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left[y_i(x^T_i \omega + \omega_0) - (1 - \xi_i)\right] - \sum_{i=1}^{N} \beta_i \xi_i.
\]

To arrive at the dual, we minimize with respect to \(\omega, \omega_0, \) and \(\xi_i\) and substitute the results from setting the respective derivatives to zero. The Lagrange dual objective function gives a lower bound on the original objective function for any feasible point:

\[
L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x^T_i x_j.
\]

The optimization problem in the dual form then can be solved by maximization with respect to \(\alpha_i\) subject to \(\sum_{i=1}^{N} \alpha_i y_i > 0\) and \(0 \leq \alpha_i \leq C, i = 1, \ldots, N\) which is simpler than solving the primal problem. The advantage is that the solution involves input features only via inner products and does not require a direct calculation of \(\omega\). The solution is a set of features with non-zero \(\alpha_i\) which are called support vectors. These features are exactly on the margin of the hyperplane and typically there are only few of them. The corresponding parameters of the primal can be obtained using these \(\alpha_i\),

\[
\omega' = \sum_{i=1}^{N} \alpha_i y_i x_i.
\]
And the prediction can be accomplished by
\[
y' = \text{sign}(f(x)) = \text{sign}(x^T \omega' + \omega_0) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i x_i^T x_i + \omega_0).
\]

The real power of SVMs is to exploit inner products when it comes to data that is not linearly separable. It can create a non-linear decision boundary by projecting data through a non-linear function \( \phi(x) \) to a higher dimensional space \( F \) where linear separation is possible. To evaluate the SVM in \( F \) rather than on the original features, every occurrence of \( x \) in the derivations above needs to be replaced by \( \phi(x) \). \( F \) is required to have an inner product \( \phi(x_i)^T \phi(x_j) \) as this is what the SVM needs to evaluate. We define the corresponding kernel \( K \) as
\[
K(x_i, x_j) = \phi(x_i)^T \phi(x_j).
\]

Since the result of the prediction only relies on the inner product of the features in \( F \) it is not necessary to perform an explicit projection to that space as long as a function \( K \) produces the same results on the original features as the inner product would yield in \( F \). This is known as the kernel trick.

### One Class Support Vector Machines

A binary SVM decides whether data belongs to either one class or another and is set up with the help of training data consisting of pairs of input \( x_i \) and expected output \( y_i \). The one class SVM in contrast predicts if new data is like the data it has been trained with. During training it is provided only the samples \( x_i \) but no expected output \( y_i \). The concept of a one class SVM has been introduced twice in literature with different derivations [Schölkopf and Williamson, 1999, Tax, 2001]. Focussing on Schölkopf’s approach, we will see how Tax and Duin arrive at a comparable solution by another line of reasoning.

Schölkopf’s one class SVM provides a function that takes value +1 in a small region capturing most of the data and −1 elsewhere. To do so, the data is mapped into a feature space by means of the kernel function. In this space the SVM separates the data from the origin and maximizes the margin. For a new point, the class membership is determined by evaluating in feature space which side of the hyperplane it falls on.

The primal optimization problem adapted to the one class case is as follows:
\[
\begin{align*}
\min_{\omega, \xi, \rho} & \quad \frac{1}{2}||\omega||^2 + \frac{1}{N} \sum_{i=1}^{N} \xi_i - \rho \\
\text{subject to} & \quad \phi(x_i)^T \omega \geq \rho - \xi_i \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots N.
\end{align*}
\]

If \( \omega \) and the offset \( \rho \) solve the problem, the decision function,
\[
y' = \text{sign}(\phi(x_i)^T \omega - \rho),
\]
will take positive values for most samples \( x_i \) contained in the training set. This can be expected because non-zero \( \xi_i \) are penalized during optimization. The parameter \( \nu \in (0, 1] \) controls the trade-off between favouring a small \( ||\omega|| \), which corresponds to a large margin, and the amount of negative classifications in the training set yielding non-zero \( \xi_i \). The one class SVM classifier
5.2 Approach

is related to density estimation which can be shown if $\nu = 1$. Then the kernel expansion of $\omega \phi(x_i)$ reduces to a Parzen window estimate of the underlying density of the data.

By solving the Lagrangian of the primal problem, the resulting dual formulation is

$$
\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j)
$$

subject to

$$
\sum_{i=1}^{N} \alpha_i = 1,
$$

$$
0 \leq \alpha_i \leq \frac{1}{\nu N}, \quad i = 1, \ldots, N.
$$

To recover the offset $\rho$, one can use the inequality constraints of the primal which in the optimum become equality constraints if $0 \leq \alpha_i \leq \frac{1}{\nu N}$. Then $\rho$ is given by

$$
\rho = \phi(x_i)^T \omega = \sum_{j=1}^{N} \alpha_j K(x_j, x_i).
$$

Notably $\rho$ is not constrained to take a positive value. For $\nu$ approaching 0, which means errors are infinitely penalized, it can take a large negative number to provide a satisfying solution to the hard margin case. A positivity constraint on $\rho$ however would cause a loss of convergence of the multipliers $\alpha_i$.

Tax and Duin [Tax and Duin, 2004] use spheres to describe the data in the feature space. The corresponding optimization problem finds the smallest hypersphere enclosing the data. The hypersphere is characterized by a radius $R > 0$ and a center $a$. The radius $R$ is the distance between the center and the boundary, and hence between the center and any support vector on the boundary. The center $a$ is a linear combination of the support vectors. The support vectors are those training points for which the Lagrange multipliers are non-zero. In order to find the smallest hypersphere enclosing the data, the volume is minimized by minimizing $R^2$:

$$
\min_{R,a} R^2 + C \sum_{i=1}^{N} \xi_i
$$

subject to

$$
||x_i - a||^2 \leq R^2 + \xi_i,
$$

$$
\xi_i \geq 0, \quad i = 1, \ldots, N.
$$

To account for outliers, a soft margin is again implemented by slack variables $\xi_i$ and a penalty parameter $C$. After solving the dual problem, the Lagrange multipliers $\alpha_i$ can be used to test if a new data point $z$ is in class or not. The new point $z$ is in class if the distance to the center is at most equal to the radius. The resolved equation reveals the similarity to Schölkopf’s algorithm:

$$
||z - x||^2 = \sum_{j=1}^{N} \alpha_j K(z, x) \geq -\frac{1}{2} R^2 + C R.
$$

It can be shown that for particular kernel functions, such as the Gaussian kernel, and data with unit norm, minimizing the hypersphere is equivalent to maximizing the margin of separation from the origin.
5 Model to Detect Horizontal Gene Transfer Into Viruses

Parameter Selection for the SVM

For prediction, we employ the one class SVM of Scikit learn [Pedregosa et al., 2011], a Python library with implementations of SVMs on the basis of libsvm [Chang and Lin, 2011]. We use a Gaussian kernel to implement a measure of similarity between two data points $x$ and $z$:

$$K(x, z) = \exp(-\frac{||x - z||^2}{2\sigma^2}).$$

As described above, the prediction result strongly depends on the choice of $\nu$. Additionally, using the Gaussian kernel requires to choose an appropriate width $\sigma$, or rather $\gamma = (2\sigma^2)^{-1}$ as it appears in the method signature of the library.

The parameter $\nu$ corresponds to the fraction of errors allowed on the training data. The training data is a set of features derived from all genes associated to a viral family. Our assumption is that there are some genes transferred into the viral family via HGT. Ideally, the choice of $\nu$ corresponds to the actual fraction of atypical genes in the data set which we cannot assess beforehand. Figure 5.1 illustrates on a toy data set how the decision boundary varies for different choices of $\nu$. To circumvent the strong dependence on the selection of $\nu$, the prediction results are reported not only in terms of $+1$ and $-1$ but as signed distances to the decision hyperplane and are sorted in ascending order. The result sorted from the most atypical to the most typical genes in a viral family of course still remains to be interpreted in terms of where to set the threshold between typical and atypical genes. It is also clear that the selection of $\nu$ not only shifts the boundary but influences its shape and hence the ordering. To obtain an ordering that is influenced by the selection of $\nu$ as little as possible, we compare the stability of the sorting for various values of $\nu$ in section 5.3.1.

$\sigma$ defines the width of the Gaussian kernel. As a rule of thumb, an appropriate choice of $\sigma$ should relate to the dimensionality of the feature space $D$ by $\sigma = \sqrt{D}$. This means the choice of $\gamma$ should be $\gamma = (2D)^{-1}$. If the width is overestimated, the exponential term will behave almost linearly and the projection to $\mathcal{F}$ will start to lose its non-linear power. If the width is underestimated, the function will lack regularization and the decision boundary will be highly sensitive to outliers in the training data. How the decision boundary varies from very wide to very narrow Gaussian kernels is depicted in figure 5.2.

The best way to select the parameters $\nu$ and $\gamma$ would be via cross validation. Since labelled data is unavailable, this unfortunately is infeasible. What remains is to evaluate the prediction quality for various choices of $\nu$ and $\gamma$.

Figure 5.1: Decision boundary of a one class SVM using different values of $\nu$. The two dimensional toy data were generated by sampling from two Gaussian distributions and adding uniformly distributed outliers.
5.2 Approach

\[ \gamma = 0.01 \]

\[ \gamma = 0.1 \]

\[ \gamma = 1 \]

\[ \gamma = 10 \]

\[ \gamma = 100 \]

Figure 5.2: Decision boundary of a one class SVM using different values of \( \gamma \). The two dimensional toy data were generated by sampling from two Gaussian distributions and adding uniformly distributed outliers.

5.2.2 Preparation of Input Features

Different input features for the one class SVM are generated from the genes of each viral family in order to compare their performance. The feature sets are derived from various genomic signatures in two different ways. Either they are based on counting statistical properties or on linear estimation of the family specific distribution of the feature.

Particular focus in this section is on the creation of codon usage estimates as inspired by metagenomics. To obtain these estimates, the frequencies of nucleotides are calculated separately for each position of a codon. The resulting characteristic set of linear equations for each viral family is then used to estimate the codon usage for a gene.

All feature sets are scaled to zero mean and unit variance in every dimension.

Avoiding Zero Probabilities

Given a data set, a probability distribution of the occurrences of certain features can be obtained by counting. If the data set is infinitely large, the derived distribution equals the true underlying distribution the data set was sampled from. As actual data sets are not infinitely large, counting only yields an approximation of the real probability distribution. For small data sets approximations are likely biased, and we might not observe some of the possible features at all.

In biological data, we generally assume that all possible features have a probability greater than zero to occur in a data set. To account for the possibility that we do not observe even a single occurrence of some feature in a sample of the underlying distribution, pseudocounts are introduced. Pseudocounts artificially elongate the data set by one occurrence for every possible feature, and thereby guarantee probabilities greater than zero for all features.

Exploiting GC Content for Linear Estimation

The GC is a widely used genome signature (compare section 2.6.1). It is the ratio of G plus C in the nucleotide sequence and is supposed to be specific for each species. We observe that the GC contents for each gene in a viral family behave linearly towards the fractions of nucleic bases in the respective genes. In agreement with Chargaff’s rule, the slopes of G and C, and the slopes of A and T coincide for families with dsDNA. Across different virus families, the slope coefficients vary. For viruses with other than dsDNA, we cannot infer the ratios of the nucleic bases only from knowing the GC content. Yet, the GC content remains an important characteristic of the genome. The linear dependency between GC content and nucleotide frequencies is shown in
5 Model to Detect Horizontal Gene Transfer Into Viruses

Figure 5.3: Linear dependence between GC content and nucleotide frequencies. The green left-hand side plots are from the Podoviridae which has dsDNA. In accordance with Chargaff’s rule, the slope coefficients of A and T, and of C and G coincide. The blue plots on the right are from Orthomyxoviridae which has a (-)ssRNA and hence do not obey Chargaff’s rule.

We use the GC content to devise linear models for each family on the basis of different features. For some features we use, the true relationship between GC content and the respective feature might actually not be linear. In fact, only for the nucleic acid frequencies, the linear dependence is motivated. The codon usage estimates we derive from the nucleic acid frequencies. For other feature sets, linear regression is used without rigorous justification. Nevertheless, we will find the approximations to provide reasonable estimates. They are capable to overcome biases exhibited in particular in short sequences because the linear regression reduces noise in the data by projection of every data point the same linear function.

Estimating Codon Usage

Different codons in a gene can yield the same amino acid in the corresponding protein. The codon usage is a $4^3$-dimensional vector with the relative frequencies of each codon in a gene. The following paragraphs in detail describe the derivation of estimates on the basis of the GC content. The reason for calculating estimates, instead of taking counts from the input sequence directly, is to overcome the bias introduced by short sequences.

Relating Nucleotide Frequencies The initial step for estimating the codon usage is to calculate the GC content of the gene as well as position based nucleotide frequencies. The latter are the frequencies of nucleotides in the three codon positions. For each gene, 13 frequencies are obtained: The frequency of A in the first position of a triplet, the frequency of A in the second, and in the third, the frequency of C in the first, second, and third position, the frequencies of T and G in all three positions of a triplet, and the GC content.

To derive information per viral species, the medians of each of the frequencies of all genes of a species are calculated. The reason to take the median and not the average is that it is less
sensitive to outliers and as we aim to identify outliers, we should expect their presence. During the evaluation, we will compare taking the median to averaging in this step. Also we consider not reducing each viral species to one vector of frequencies, but to continue the algorithm with nucleotide frequencies per gene.

From the nucleotide frequencies per species (respectively per gene) in a family, a statistical model is constructed for each viral family. As observed in an earlier study frequencies of nucleotides in the three codon positions depend linearly on global nucleotide frequencies [Besemer and Borodovsky, 1999]. In turn, this means nucleotide frequencies in the three codon positions depend linearly on the GC content. The slope coefficients however are distinctly different for each frequency. Taking the values of all species in a family, a linear regression is calculated for each frequency of nucleotides in relation to the GC content. The result is a set of twelve linear equations for each viral family. A plot of the position based nucleotide frequencies and the linear functions obtained by regression is shown by the example of *Podoviridae* in figure 5.4

**Estimate of Codon Usage** We estimate the codon usage following the procedure suggested by Zhu and co-workers [Zhu et al., 2010] for metagenomic samples. Because of the strong correlation of the global nucleotide frequency with the genome GC content, the constructed set of
linear equations can be used to calculate the codon usage of a viral gene. Given the GC content of a viral gene, the linear equations of its family provide position based nucleotide frequencies. We assume that the frequency of a codon is proportional to the product of the respective position based frequencies. More formally, let us denote the position-based frequencies by \( f_{Xk} \) and the codon frequency by \( f_{XYZ} \), where \( X, Y, Z \in \{A, T, C, G\} \) and \( k \in \{1, 2, 3\} \). Then we assume that

\[
f_{X1} f_{Y2} f_{Z3} \sim f_{XYZ}.
\]

This proportionality yields an initial approximation of codon frequency \( f'_{XYZ} = f_{X1} f_{Y2} f_{Z3} \). Additional correction comes from the frequency of the encoded amino acid \( \alpha \) and the number of synonymous codons is taken into account as well. The estimate of codon frequency by the example of GCT, which is one out of four synonymous codons that encode for alanine, is then

\[
f_{GCT} = f_{\text{alanine}} \times \frac{f'_{GCT}}{f'_{GCT} + f'_{GCC} + f'_{GCA}}.
\]

The frequency of the encoded amino acid \( f_{\alpha} \) is derived by linear regression on the GC content of all frequencies of that amino acid in that family.

The codon frequency estimates are calculated for every codon. Each gene then is represented by 61 values, which is the number of all possible codons minus the stop codons. Stop codons do not encode for amino acids and are excluded from the database records of coding sequences we work with. The 61-dimensional vector with the frequencies relative to the total amount of codons serves as input set for the one class SVM.

**Employed Feature Sets**

The feature sets evaluated for their ability to serve as input to the one class SVM can be grouped by two criteria. Distinguishing the method of derivation of the feature sets, we compare sets derived by linear estimation and sets derived by counting. Distinguishing the type of signature, the derivation of the compared feature sets either involves triplet frequencies or word counts.

The feature sets involving triplet frequencies are:

- **Amino acid frequencies**
  
  Amino acid frequencies are the relative amounts of all amino acids in the translation of a gene. Counting them results in a 20-dimensional feature vector.

  For the estimate version, for each viral family, the frequencies of each amino acid in each protein are plotted against the GC content of the corresponding gene. For each amino acid, a linear function is fitted. Provided the GC content of a gene, the evaluation of these functions yields a 20-dimensional vector of amino acid frequency estimates.

- **Codon usage**

  The frequencies of each codon relative to the total amount of codons yield a vector with 64 dimensions of which we use the 61 dimensions that do not refer to stop codons as features.

  To obtain estimates of codon usage, the frequencies of nucleotides are calculated separately for each position of a codon. Regression on the GC content results in a characteristic set of linear equations for each viral family. As described in detail in the previous section, this set is then used to estimate the 61-dimensional vector of codon usage for a gene.
5.2 Approach

- **Relative codon usage**
  The frequencies of each codon relative to its synonymous codons neglecting the stop codons yield a 61-dimensional feature vector.
  The estimated version is derived from the estimates of codon usage. The relative frequencies of synonymous codons form a 61-dimensional feature vector.

- **Position based nucleotide frequencies**
  The frequencies of nucleotides in the three codon positions, as used to derive a codon usage estimate, yield a 12-dimensional feature vector. This rather primitive feature set is included for the sake of comparison only.

Of these feature sets, we evaluate also variants. One variant is to amend each feature vector by the GC content. The other variation affects the estimation procedure for the first three feature sets. For estimated amino acid frequencies, estimated codon usage, and estimated relative codon usage the variants are: taking the median or taking the average of the position based nucleotide frequencies per species, or proceeding to the next step, namely the linear regression, directly. The last variant implies that the linear regression is calculated over the position based nucleotide frequencies of all genes of a viral family, and not over the position based nucleotide frequencies of all species of a viral family as in the other two cases.

The other type of feature sets are constructed from the nucleotide sequences on the basis of word counts. These feature sets are:

- **Oligonucleotide frequencies**
  Oligonucleotide frequencies are the frequencies of words of a defined length in a sequence. We create feature sets each for length one, two, three, and four. While the feature vector for length one only has four entries resembling the nucleotide frequencies, the feature vector for length four has $4^4 = 256$ entries. We do not consider longer k-mers, because the counts for certain words are too low and potentially biased, especially in smaller virus families.
  To obtain an estimated version of each of these feature sets, linear regression of oligonucleotide frequencies on the GC content is conducted.

- **GC content**
  The very simple feature set consisting only of the GC contents of each gene is used in order to verify that the one class SVM model with sophisticated features as input is of actual use. The GC content is not supplied as to the SVM but a very naïve classifier is constructed instead. This simple classifier calculates the median GC content of all genes in a viral family and ranks each gene by its distance to the median. Thus, the genes furthest away from the median are the most atypical ones.

5.2.3 Measures for Validation

To evaluate the performance of the prediction algorithm, several analyses are carried out. Due to the lack of ground truth data, there is no direct indicator of quality of the prediction results. We have implemented two validation techniques. One validation approach is to run the prediction on simulated data sets. These data sets are formed by all genes of a virus family plus a defined proportion of alien genes chosen at random from all virus genes not from this family. The
other approach is to interpret the prediction result with the help of a BLAST query for each gene. To assess the performance of the unsupervised prediction algorithm, we make use of various statistical measures which are described below. As part of the evaluation is carried out on simulated data, for which we know the labels, some measures are specific to labelled data.

**Spearman’s Rank Correlation**  Spearman’s rank correlation coefficient, also denoted as Spearman’s \( \rho \), measures the correspondence between rankings. It is a non-parametric measure of statistical dependence between two variables, calculated by using the ranks in place of the actual observations in the formula for the correlation coefficient [Zwillinger and Kokoska, 2010].

Given \( n \) pairs of observations from two continuous distributions, the observations in the samples are ranked separately from the smallest to the largest. Equal observations are assigned the mean rank of their positions. With \( u_i \) being the rank of the \( i \)-th observation in the first sample and \( v_i \) being the rank of the \( i \)-th observation in the second sample the correlation coefficient \( \rho \) is calculated as

\[
\rho = \frac{\sum_i (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_i (u_i - \bar{u})^2 \sum_i (v_i - \bar{v})^2}}.
\]

It rates how well the dependence between two variables can be described by means of a monotonic function. Like other correlation measures, \( \rho \) can take values between \(-1\) and 1. Correlations of \(-1\) or \(+1\) imply an exact monotonic relationship, 0 indicates that there is no correlation between the variables.

A hypothesis test with the null hypothesis that the two sets of observations are uncorrelated reveals the significance of the measured correlation. The test yields a p-value indicating the probability of an uncorrelated system to produce datasets that have a Spearman correlation like the one observed or better.

We use this measure to assess how comparable results of the predictor under varying parameter sets are. The quality of the prediction is higher if the result is invariant under small perturbations of the parameters. If the ranking varies a lot with only slight changes of the parameters, the prediction is unstable and thus not likely to provide a correct solution.

**Matthew’s Correlation Coefficient**  Matthew’s correlation coefficient (MCC) is a measure of the quality of binary classifications. To describe the performance of a binary classification, typically the amounts of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn) are provided. These values refer to the number of predictions classified correctly as positive, classified wrongly as positive, classified correctly as negative, and classified wrongly as negative. Hence, the sum of tp plus fn is the actual number of positive samples, and the sum of fp plus tn is the actual number of negative samples in the data set. The four values together provide a complete description of the classification performance of an algorithm but this description lacks conciseness compared to a single value indicating performance. MCC combines tp, fp, tn, and fn into a single score.

Unlike other measures, such as sensitivity (\( \frac{tp}{tp+fn} \)) or specificity (\( \frac{tn}{fp+tn} \)), MCC takes into account tp, fp, tn, and fn. An advantage over accuracy, which also takes into account all four measures and is calculated as \( \frac{tp+tn}{tp+tn+fp+fn} \), is that it can be used even if the classes are of very different sizes. MCC ranges from \(-1\) to 1 and is defined as

\[
\frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(m + fp)(m + fn)}}
\]
5.2 Approach

Figure 5.5: Behaviour of MCC scores for different class proportions, depicted in different colors, under variation of the percentage of errors within both (left), only the positive (middle), and only the negative class (right).

A score of 1 refers to a perfect prediction while −1 indicates perfect negative correlation. A score of 0 corresponds to a random distribution [Matthews, 1975].

Because the data in which to predict atypical genes is probably highly imbalanced, one has to understand how the MCC behaves in case of imbalanced data. Figure 5.5 shows MCC scores given a percentage of errors among both, the positive and the negative samples, as well as scores if only errors in the positive or in the negative samples occur. Each line in the plot corresponds to a given ratio between the amounts of positive and negative samples in the data set.

Notably, in case of highly imbalanced data, although the range of the scores remains unchanged, the behaviour is very different from linear. If equal amounts of errors are made among the positive and the negative samples, a score much lower than in the balanced case corresponds to equally good prediction quality. The same is true if the errors only occur in the larger, here the positive class. On the contrary, if errors only occur among the negative samples, which is the much smaller class in this case, low MCC scores are even worse than in the case of balanced classes.

To judge if a medium MCC score is satisfying, we compare MCC and specificity. The specificity only considers negative samples and behaves linearly irrespective of the ratio between the classes. It yields a value of 0 if no negative samples were predicted correctly and a value of 1 if all negative samples were predicted correctly. Sensitivity could be considered likewise, if the positive class would be the small one. In the evaluation, we require the MCC score to be lower than the specificity for a high quality prediction.

**Statistical Significance** To test whether the observed MCC scores are significant, the prediction is carried out also on randomly labelled data. This means, the labels indicating if a gene is an artificially introduced foreign gene are distributed at random over the data set. The MCC score then is calculated with respect to the random labels. With many repetitions of the same procedure, \( N = 100 \) in our case, we can estimate the standard deviation (SD) of the MCC scores.

\[
\text{SD} = \sqrt{\frac{\sum (a_i - \bar{a})^2}{N - 1}}
\]
The SD indicates how much the MCC scores $a_i$ vary from their mean $\bar{a} = 0$. With the scores being normally distributed, 68.2% of the scores fall between $±1$SD and $−1$SD, and 95.4% of the scores fall between $±2$SD and $−2$SD. This means, if we observe a score larger than $±2$SD or smaller than $−2$SD, it is significantly different from random with a significance level of 5%.

**Prediction Success** To evaluate the prediction algorithm for the different sets of input features and different SVM parameters $\nu$ and $\gamma$ we establish a measure of success. The prediction of atypical genes in a viral family is considered successful if the following conditions hold:

- The MCC score is statistically significant (at significance level of 5%).
- The MCC score is smaller than the specificity.
- The MCC score is greater than 0.4.

In case the other conditions hold but the MCC score is only greater than 0.2, we call the prediction semi-successful. The prediction for a family is considered successful if success is reported at least once for one combination of feature set and parameters.

**Cluster Silhouette** The silhouette value is a measure of cluster goodness first described by Peter Rousseeuw [Rousseeuw, 1987]. It is applicable to the result of any clustering technique and needs no other information than the clustered data itself. It captures if the cluster members are close to each other and far away from members of other clusters.

The silhouette $s$ value of the $i$-th sample in the data is defined as

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

where $a(i)$ denotes the average dissimilarity of sample $i$ with all samples in the same cluster, and $b(i)$ denotes the lowest average dissimilarity of sample $i$ with all samples from another cluster. In other words, $a(i)$ measures how well sample $i$ fits in the cluster it got assigned to, and $b(i)$ measures how well sample $i$ fits the neighbouring clusters. The value of the dissimilarity measure is lower the better the sample matches the other samples in the cluster. Commonly, a distance measure is used to calculate dissimilarity. Values for $s(i)$ range between $−1$ and $+1$. The silhouette value of a cluster is the average over the $s(i)$ of the cluster members, and the silhouette value of a whole clustering is the average over all $s(i)$. For a clustering, an average silhouette value greater than 0.5 indicates reasonable partitioning of data. A value smaller than $−0.2$ means that the data do not exhibit cluster structure [Kaufman and Rousseeuw, 1990].

We only consider the genes identified as typical for the calculation of the silhouette value, as the central feature of the one class SVM is to distinguish samples alike one another from dissimilar samples. The dissimilar samples are heterogeneous and cannot be regarded as a cluster.

The reason to include silhouette values in the evaluation is that they can be calculated for both, the real and the simulated data sets. By showing the similarity of both sets of attained silhouette values, we indicate that the performance of the predictors is comparable when trained with real and simulated data sets.
5.2 Approach

Area Under Curve  To get an idea how good a ranked prediction result is, enrichment plots are evaluated. These plots show the cumulative sum of all successful recoveries of atypical genes plotted against the rank in the result.

In the ideal case, all atypical genes are found first and then the typical genes are enumerated. As each atypical gene adds 1 to the cumulative sum and the typical genes add nothing, the plotted curve will ascent steeply and then remain constant at its maximum. In the worst case, the ranking is inverse: First the predictor reports all typical genes, then the atypical ones. The corresponding curve would first remain constantly zero and then steeply ascent reaching its maximum in the last position of the ranking. A random prediction result would be close to the diagonal line in this plot indicating no preference for typical or atypical genes.

The normalized version of this plot is called receiver operator characteristic (ROC) and depicts the true positives rate plotted against the false positives rate. It contains the same information as the plot of cumulative sums but the data is scaled to proportions between 0 and 1.

The prediction quality measure derived from the ROC plot is the area under curve (AUC). This is the proportion of the area between best case and worst case behaviour that is covered by a curve of the prediction result. The use of AUC as a measure of quality for predictions is debated, as the comparison between two ROC curves is only meaningful if the curves do not cross. Yet, the evaluation of ROC curves is suitable for our purposes, because they show the ranking potential of a method. The ranking potential is related but not equal to the overall performance [Vihinen, 2012], and our primary interest is the ranking induced by the signed distances of the genes to the decision boundary of the SVM.

For the interpretation of AUC scores, Swets suggested guidelines: Scores higher than 0.9 indicate excellent performance, scores of 0.8-0.9 are good, scores of 0.7-0.8 are fair, scores of 0.6-0.7 are poor, and scores lower than 0.6 indicate fail [Swets, 1988].

We use this measure for both the simulated data sets and the real data. Clearly, for the simulated data sets, atypical genes are the ones artificially added. The real data, we classify based on a BLAST search. If a gene has a BLAST hit outside its viral family, and thus there exists a possibility that it has been introduced into the viral family via HGT, it is considered atypical. This labelling is at most an approximation to the truth, but ideally should reveal a trend. Many genes with BLAST hits outside the family may have been inherited from a common ancestor.

5.2.4 Steps of Evaluation

The assessment of prediction performance involves the aforementioned variety of measures and comprises multiple analyses. In this section we provide an overview of the different steps of performance analysis.

Some of the performance measures require knowledge about true labels. Therefore we generate simulated data sets. For each viral family, the set of genes is extended by a small proportion of genes not from that family. We assume that the method established to predict atypical genes is as capable to recover these foreign genes. Importantly, for the analyses on simulated data, the models are already trained on the simulated data sets. As it is not possible to exclude the true outliers from the real training set during training, likewise the simulated outliers are not excluded. To keep the influence of the additional data low during training, the amount of added simulated outliers does not exceed 5%.

Overall, the evaluation comprises the following steps:
5 Model to Detect Horizontal Gene Transfer IntoViruses

1. Analysis of ranking stability

A valid prediction result should be robust towards small changes of parameters. We analyze the similarity of the ranked output of the one class SVM under varying parameters $\nu$ for several values of $\gamma$ and different feature sets. We aim to select $\nu$ and $\gamma$ such that the prediction result becomes stably ranked. To compare rankings, Spearman’s rank correlation coefficient is evaluated.

2. Measuring prediction performance on simulated data

The robustness of the prediction result is no indication for its correctness. To assess how well alien genes can be discovered, we evaluate for different feature sets how well the simulated alien genes become recovered employing two different measures, the AUC and the MCC scores. For the MCC scores we assess also their significance to decide on the success of a prediction. Obviously, we cannot expect perfect recovery rates for the simulated alien genes as the actual outliers remain unlabelled in the simulated data set and compete for the highest ranks in the result.

3. Comparability to real data sets

To judge whether the prediction performance on real data is comparable to that on simulated data where labels are at hand, we compare the behaviour by means of the silhouette value. The shape of the boundary between typical and atypical genes is expected to be similar irrespective of the presence or absence of the simulated data. Correlation in the silhouette values for real and simulated data sets can confirm this similarity.

4. Performance on real data

Performance of the prediction method on the real data is assessed by measuring the AUC inferring labels from BLAST searches for each gene in the result. Genes are considered atypical if they have at least one BLAST hit outside the viral family. As this labelling can at most reveal a trend, a literature search is conducted for a number of genes to strengthen the analysis.

5.3 Results

We evaluate the capability of the one class SVM to discover atypical genes under varying parameters and with different feature sets. The data are all viral genes from ENA, downloaded in July 2013 and pruned for redundancies using CD-HIT [Li and Godzik, 2006] with a sequence similarity level of 95%. CD-HIT clusters sequences with the defined sequence similarity together and returns one representative for each cluster. The representatives, grouped by their virus family, yield the data sets. The data sets form the basis to derive feature sets presented as input to the prediction algorithm. The expected result of the one class SVM, given a feature set and parameters $\nu$ and $\gamma$, is a ranked list, sorted from the most atypical gene to the most typical gene in the respective virus family.
5.3 Results

Figure 5.6: Correlation between prediction results under varying $\nu$ for different values of $\gamma$ (corresponding to the different colors). The prediction results originate from *Partitiviridae* using codon usage estimates in the upper plot and relative codon usage without regression in the lower plot. Every point corresponds to the correlation between the prediction result for one value of $\nu$ and the prediction result for the next value of $\nu$, averaged with the correlation between the result for current value of $\nu$ and the previous. The step size between subsequent values of $\nu$ is 0.001.
5 Model to Detect Horizontal Gene Transfer Into Viruses

5.3.1 Robustness of the Prediction

Robustness of a prediction result towards small perturbations of the parameters is one indication of quality. A stable result is not necessarily correct but more trustworthy than a result that immensely varies under small disturbances. To assess the robustness, we measure the stability of the prediction result under variation of $\nu$ for every feature set and different values of $\gamma$. A prediction result is stable if there is no or only little difference in the ranked list of genes, obtained from the one class SVM with one or another slightly modified parameter set. The slight modification of the parameter set is attained by variation of $\nu$ while keeping the other parameters constant. The parameter $\nu$ takes 1,000 equidistant values in the interval $(0, 1]$. The similarity of the ranked lists, obtained for successive values of $\nu$, is measured by Spearman’s $\rho$.

Figure 5.6 illustrates our findings by the example of *Partitiviridae* for two different parameter sets, codon usage with linear regression and relative codon usage without regression. Each point in the plots depicts the average Spearman correlation coefficient over the prediction result with the current value of $\nu$ and the previous, and the prediction result with the current value of $\nu$ and the next. A value of 1 indicates that the ranking with the current value of $\nu$ is equal to the ranking resulting from the preceding value of $\nu$ and the subsequent value of $\nu$.

For codon usage estimates, the feature set used for the upper plot, we observe that for high $\nu$, larger than 0.8, the ranking remains invariant under changes of $\nu$ irrespective of the choice of $\gamma$. The smaller $\nu$ is, the less correlation between neighbouring parameter sets is observable. Additionally, the stability depends on $\gamma$. For small values of $\gamma$, the stability is generally better than for large values of $\gamma$. As shown for the relative codon usage feature set, in figure 5.6 at the bottom, the prediction result obtained with large values of $\gamma$ can be very sensitive to small deviations in $\nu$. For this feature set, if the value of $\gamma$ is large, there is not even a trend towards increasing stability when $\nu$ is large.

For all parameter sets and for all virus families we observe that $\gamma = (2D)^{-1}$ yields a high stability over a large range of $\nu$. This finding confirms that the rule of thumb for the selection of $\gamma$ is a reasonable choice. Thus, we refrain from conducting the subsequent evaluation for multiple values of $\gamma$ and fix it to $\gamma = (2D)^{-1}$, where $D$ is the dimensionality of the respective feature space.

The parameter $\nu$ is supposed to correspond to the proportion of outliers we expect to observe in the data. This proportion is unknown and can only be guessed. The ranking stability analysis indicates that assuming the proportion of outliers to be very low, for instance one percent, might yield an unstable prediction result. Taking into account not only the binary labels from the SVM result but the signed distances to the hyperplane to obtain a ranking, we can fix $\nu$ such that the ranking is stable irrespective of assumptions on the actual number of alien genes. To find a reasonable value for the parameter $\nu$, we consider the results obtained on simulated data sets for various choices of $\nu$.

5.3.2 Prediction Quality on Simulated Data

The quality of the prediction result on simulated data sets is assessed using the measure of success established in section 5.2.3 on the basis of MCC. Simulated data sets consist of all genes of a viral family plus 5% of random alien genes from other viruses. Consequently, in order to measure MCC, 5% of the genes with the lowest signed distances from the decision boundary are considered as predicted atypical genes. With this measure we choose an appropriate $\nu$ and
characterize the feature sets. Furthermore, we assess AUC scores which we also survey for the evaluation on real data sets.

93 families have more than three genes and hence are large enough to run the prediction on. For 77 of them, our method is capable of identifying the alien genes with statistically significant MCC larger than 0.2 in at least one set of features and for at least one choice of \( \nu \), denoted semi-success. 62 of the families fulfil the criteria for success, meaning that the MCC is larger than 0.4. The number of genes in the families ranges from less than 100 for various families to more than 10,000 for Phycodnaviridae and Flaviviridae. The family of Retroviridae is too large to be processed at once as it comprises almost all retroviruses. Therefore this family is split up taking a lower level of taxonomy, the level of genus. On this level, the genus Lentivirus still comprises more than 100,000 genes from which we sample only 25,000.

The large deviations in the number of genes per family are the result of extensive study of some viruses. In numerous studies many strains of them have been sequenced and included in the databases. The family of Flaviviridae for example comprises viruses such as Dengue virus, Yellow fever virus, Hepatitis C virus, and Swine fever virus. These viruses are infectious to humans or domestic animals and are therefore of particular interest. Lentivirus includes HIV which is subject of intensive research as it causes the acquired immunodeficiency syndrome (AIDS). Phycodnaviridae infect algae and are probably so numerous in the database because of metagenomic studies of water samples.

The prediction success of the one class SVM does not depend on the size of the viral family. We observe success rates between 55% and 83% with no clear trend regarding the family sizes (compare figure 5.7). The semi-success rates are constantly above 80% except for the families with more than 4,000 genes. This might hint on a decreased predictive power for large families because they already exhibit large variances with their own genes and might conceal actual outliers within it.

![Figure 5.7: Distribution of the number of genes per family (blue) and the success rates in each bin (semi-success lightgreen, success darkgreen).](image-url)
Figure 5.8: Success rates of the one class SVM for different values of $\nu$. The proportion of successfully predicted families is depicted in darkgreen, the proportion of semi-successes in lightgreen.

**Choice of $\nu$**  
Assessment of the prediction success for various values of $\nu$ supports the presumption obtained during the analysis of the robustness of the classification result in section 5.3.1. Very small $\nu$, smaller than 0.1, yield half as many successes as larger $\nu$. The semi-success rates constantly are around 80%. Among the larger $\nu$s, there is no preference for a specific value. The prediction success in relation to $\nu$ is visualized in figure 5.8. We fix $\nu = 0.2$ for the subsequent analysis.

**Comparison of Feature Sets**  
A comparison of prediction performance using the different feature sets described in section 5.2.2 is conducted by means of the number of successfully predicted families. As the measure of success, although well motivated, uses arbitrarily chosen thresholds (see section 5.2.3), the amount of successes does not indicate the absolute prediction performance. However, it allows to compare the performances of the different feature sets against each other.

Success rates for the different feature sets vary between 4% (15% semi-success) for the GC content and 48% (63% semi-success) for tetranucleotide frequencies without regression as depicted in figure 5.9. All feature sets derived by estimation through linear regression on the GC content perform equally well. Their success rate is 30% (60% semi-success) irrespective of whether the derivation involves triplet frequencies or oligonucleotide frequencies or whether the feature set is amended by the GC content. Also the variation in the estimation procedure (the use of the mean or the median of the values for genes of a species or neither) yields no difference in the success rates. While this variation is expected to have minor influence, the reason for all the different signatures to perform equally when derived by linear estimation is not obvious and has its origin in the standardization procedure applied before supplying the data to the SVM.

Scaling the data to zero mean and unit variance transforms the lines obtained from linear regression such that they go through the origin and have a slope of either $+1$ or $-1$. This means, irrespective of the dimensionality of the feature set, the number of distinct values in the feature
5.3 Results

- est. codon usage (avg. per gene)
- est. relative codon usage (avg. per gene)
- est. amino acid frequency (avg. per gene)
- est. codon usage + GC (avg. per gene)
- est. relative codon usage + GC (avg. per gene)
- est. amino acid frequency + GC (avg. per gene)
- est. codon usage (median)
- est. relative codon usage (median)
- est. amino acid frequency (median)
- est. codon usage + GC (median)
- est. relative codon usage + GC (median)
- est. amino acid frequency + GC (median)
- codon usage
- amino acid frequency
- codon usage + GC
- relative codon usage + GC
- amino acid frequency + GC
- mononucleotide frequency
- dinucleotide frequency
- trinucleotide frequency
- tetranucleotide frequency
- est. mononucleotide frequency
- est. dinucleotide frequency
- est. trinucleotide frequency
- est. tetranucleotide frequency
- position based nucleotide frequency
- position based nucleotide frequency + GC
- est. position based nucleotide frequency
- GC

Figure 5.9: Success rates of the one-class SVM for different feature sets. The proportion of successfully predicted families is depicted in dark green, the proportion of semi-successes in light green.
vector is always two if we apply linear regression and data scaling. Spearman’s rank correlation coefficient confirms that not only the amount of successes but also in which families they occur is equal for all feature sets when using linear regression.

The success rates of the feature sets based on counts exhibit differences. While we observe success rates lower than those of the estimation based feature sets for amino acid frequencies, position based nucleotide frequencies, and for the GC content, the codon usage, the relative codon usage, and the oligonucleotide frequencies outperform the estimation based feature sets. The best score of the triplet based feature sets, 40%, is exhibited by the relative codon usage without regression feature set. This feature set, if amended by the GC content, even improves to 46% (66% semi-success). A similar improvement when adding the GC content we also notice for amino acid frequencies and codon usage. The advance is well in agreement with the success rate of predicting solely on the GC content. In their success rate, tetranucleotide frequencies without regression exceed all other feature sets and yield a success rate of 48%. The semi-success rate of 63% however is a bit lower than that of relative codon usage plus GC content without regression.

With the estimation using linear regression, our objective is to overcome the potential statistical bias introduced by the rather short genomes of viruses displaying high mutation rates and extreme diversity. The linear regression on the GC content is a projection onto the axis of the largest variance. It removes degrees of freedom from the data with the intend of concealing noise. For the amino acid frequencies this projection yields an increased performance indicating that insufficient statistical information have strong negative impact on the result for the amino acid frequencies feature set. While without linear regression, the success rate is 17% (45% semi-success), the estimated amino acid frequencies yield 30% (60% semi-success). The opposite we observe for codon usage, relative codon usage, and oligonucleotide frequencies. The linear regression appears to have bad effect on the predictive power of these feature sets. It conceals not only noise but also critical information.

**Impact of Standardisation** The feature sets are standardized to zero mean and unit variance before submitted to the SVM. We compare the results obtained when omitting this step to the findings described above. The reason for this investigation is the de facto employment of only two distinct features when using linear regression plus standardization.

When using no standardization, the success rates of the feature sets derived by linear regression are slightly different from one another, however all in the range between 32% and 36%. The best success rates, 36%, we observe for average codon usage and estimated di- and tri-nucleotide frequencies. The feature sets without linear regression show success rates of 40% for the oligonucleotide frequencies of length three and four, also above 36% for the ones of length one and two, and below 32% for the triplet based feature sets. The semi-success rates are stable around 60% for all feature sets except amino acid frequencies without regression which yield only 35%. Hence, compared to the standardized feature sets we observe a decrease in the success rates of the feature sets derived without linear regression and a slight increase in the success rates of the feature sets derived with linear regression.

Desisting from scaling allows for more variation in the slopes of the linearly regressed features. These deviations in slopes explain the variation in the success rates. Unlike in the standardized setting, we observe as many different slopes as there are dimensions in the feature set. However, due to the lack of standardization, extremer slopes probably have more impact on the prediction than slopes close to zero, and thus the number of dimensions important to the
5.3 Results

prediction might again be fairly small. Not standardized, and therefore variable, slopes are less successful compared to the feature sets without regression that exhibit the same freedom but yield success rates up to almost 50% in the standardized setting.

Another observation is that the prediction success is invariant under different choices of $\nu$. As in the standardized case, the semi-success rate for all $\nu$ is about 80%. In contrast to the standardized case, the success rate is also constant at 60% for all choices of $\nu$. Thus albeit the ranking of predictions may vary, the choice of the atypical genes stays the same for different values of $\nu$.

**GC Content Range**  Asking whether there is any pattern behind virus families being successfully predicted, we relate the prediction success to the range of GC content the virus family exhibits. GC content levels from genes of the same family (including the simulated atypical genes) may vary by 75%. Most families however exhibit deviations below 40%. As figure 5.10 shows, the prediction success is higher than 60%, and higher than 80% including semi-successes, for GC content ranges between 15% and 47%. On very narrow ranges, 8-15%, the prediction is successful in more than half of the cases. Typically, such narrow ranges are observed if there are only few genes of a family in the database.

For the few examples of wide ranges of GC content, greater than 49%, the prediction is semi-successful in only one case. These families can be suspected to also show large deviations in the other recorded statistics and thus the outliers might be hidden within the diversity of the family. As the simulated outliers are chosen at random, there is no guarantee that they reside outside the ranges of statistics typical for the respective family. On the contrary, they are likely to be concealed if the deviations within the family are large.

![Graph showing distribution of GC content ranges per family and success rates](image)

Figure 5.10: Distribution of GC content ranges per family (blue) and the success rates (dark-green) and semi-success rates (light-green) in each bin.
Prediction Success in Different Viral Classes  Evaluating the prediction success with respect to the viral classes we notice that the prediction is not equally successful for each viral class. As shown in figure 5.11, the prediction success is overall good for retroviruses, excluding the prediction solely based on GC content. Most feature sets yield comparably good performance on (+)ssRNA and (-)ssRNA viruses. The feature sets based on linear regression perform well for (+)ssRNA but not for (-)ssRNA. Position based nucleotide frequencies appear to be less capable for (+)ssRNA. Relative codon usage plus GC content is the only feature set that exhibits comparatively high success rates for dsRNA viruses. The success rates for the prediction for dsDNA viruses are comparatively low on all feature sets. Tetranucleotide frequencies appears to be the only feature set suited for ssDNA viruses.

The simulated outliers are chosen at random out of all viral genes except the genes from the family they are added to. As the families are of different sizes, the selection of outliers is biased towards families with many genes, and hence towards virus types many genes stem from. Half of all genes stem from retroviruses and approximately 20% each from dsDNA and (+)ssRNA viruses, so these classes can be expected to perform worse in the test, agreeing with the observation for dsDNA viruses. However, our method performs extremely well for retroviruses, which in turn may be explained by the fact that some genera of retroviruses (e.g. Lentivirus) are much more populated than others.

![Figure 5.11: Prediction success per virus type on different feature sets. Semi-success rates are omitted. The colors indicate the virus type. Numbers in braces denote the number of families belonging to that type.](image-url)
5.3 Results

Figure 5.12: Boxplot of AUC scores on different feature sets. Red bars indicate the median score observed in all families and 50% of all scores fall within the range indicated by the blue boxes. The horizontal line at 0.5 highlights the score expected for a random result.

**Ranking Potential**  One might argue that we established the measure of success just such that it works well. An evaluation of ROC curves approves the results found, studying the measure of success based on MCC. As depicted in figure 5.12, the AUC scores for the different feature sets exhibit the similar trends as observed for the success rates (figure 5.9). With the success rates however, we could not judge on the absolute quality of the predictions. According to the AUC score classification system by Swets [Swets, 1988], the one class SVM on median yields good and excellent rankings for the triplet based feature sets without regression and the oligonucleotide frequencies without regression. This means, we observe median AUC scores between 0.82 and 0.92. The feature sets based on linear regression yield a fair median score of 0.72, and the median score for the GC content is in the range of fail with a value of 0.59.

Although comparison between different scores might be inappropriate because crossing of the ROC curves is not precluded, we want to mention that, unlike in the analysis of success rates, the AUC scores for amino acid frequencies without regression and position based nucleotide frequencies are not lower than the AUC scores for feature sets using linear regression. The opposite we observed in the analysis of success rates depicted in figure 5.9.
Figure 5.13: Boxplot of silhouette values on different feature sets. The top plot shows the scores for prediction on data with simulated atypical genes and the bottom plot for the real data sets. The silhouette value for GC content is not depicted as it has not been recorded.
5.3 Results

5.3.3 Comparison of Ranking Potential on Real and Simulated Data

There is no direct way to prove that the prediction on the real data sets is as successful as observed for the data sets with simulated outliers. To provide evidence that the behaviour is similar, we compare the silhouette values on both types of data sets. They indicate cluster goodness and are expected to show similar trends on the data with and without simulated outliers.

We observe that the silhouette values overall are not notably high. The best median values are 0.66 on the simulated data and 0.56 on the real data. Both these values are observed for the estimation based feature sets and indicate reasonable partitioning of the data. Median silhouette values for the other feature sets are roughly in the interval between 0.2 and 0.5, which is the zone between no partitioning at all and reasonable partitioning. The scores on simulated data are in each case by about 10% higher than on the true data but exhibit similar behaviour, confirmed by a reasonable Spearman coefficient of 0.65 with a p-value beyond float precision: The higher the dimensionality of the feature set, the lower the observed silhouette value. Feature sets derived by linear regression are de facto two-dimensional due to the standardization procedure and yield the best silhouette values. Feature sets of tetrancleotide frequencies are 256-dimensional and yield the lowest observed silhouette values of 0.24 on the simulated data and 0.16 on the real data. Figure 5.13 shows the silhouette values for selected feature sets on both data sets in comparison.

The observed dependence of the silhouette values on the feature set dimensionality is explained by the curse of dimensionality. One can show that distance measures in high dimensional space are not meaningful and are sensitive to the metric chosen [Aggarwal et al., 2001]. The silhouette values have been obtained incorporating the euclidean distance metric. When the dimensionality of the feature space increases, distances to the nearest neighbour approach the distance to the farthest neighbour. This means, the contrast in differences to different data points becomes non-existent [Beyer et al., 1999]. Beyer and co-workers demonstrate that this effect already impairs results on 10-dimensional data.

The silhouette values indicate similar behaviour of the real and the simulated data. As for their apparent dependence primarily on the dimensionality, these scores might not be suited to compare the goodness of clustering for feature sets of different dimensionality. Worth consideration might be whether the poor silhouette values are an indication of overfitting. However, apart from the concerns about the meaningfulness of the silhouette values, the confirmed robustness of the prediction result speaks against this.

5.3.4 Prediction Quality on Real Data

We conduct a second evaluation to obtain an indication of quality of the prediction on real data. The prediction results are compared with BLAST search results for every gene in the ranked result of the SVM. The BLAST search is executed on protein level using version 2.2.25 with default parameters. Default parameters implies that each result comprises at most 500 hits and the maximum expect value for each hit is 10 which allows to include distant results at the cost of confidence concerning the relevance of each hit.

Assuming a gene is atypical for a family when it has BLAST hits outside the family, we derive a labelling for the genes. This labelling at most provides a tendency and is clearly inaccurate. Besides biases from unequally well studied organisms, it does not differentiate between hits obtained outside a family. A BLAST hit outside the virus family can occur due to gene transfer from the family to another, or due to gene transfer into the family, or simply may be a signature
of a common ancestry. While with this naïve labelling, all three cases are classified atypical, the prediction algorithm is supposed to identify only the second case. The observation of concordance between the prediction result and the labelling derived from the BLAST searches, despite these shortcomings, provides support, yet no proof, for the established methodology. On the contrary, inconsistency between the labelling and the prediction result does not imply failure of the prediction algorithm. In any case, the most valuable support, if at all, has to be provided by extensive literature search.

BLAST searches despite their reasonable performance become computationally expensive when conducted for each gene in a family. Therefore, we restrict this analysis to 22 families. Considering the distributions of scores from all these families at once, as depicted in figure 5.14, yields no indication of success for any of the feature sets. All median scores are close to 0.5. Overall, we thus cannot provide evidence whether the prediction algorithm is successful using the labelling inferred from BLAST searches for each gene.
5.3 Results

Orthomyxoviridae  However, for some families we observe the prediction result and the labelling inferred from BLAST searches to be partially in conjunction with each other. This is manifested by AUC scores considerably higher than 0.5. The most distinct such family is Orthomyxoviridae. The ROC curves in figure 5.15 show the resemblance of the SVM result and the BLAST search based labelling. The AUC scores of the best performing feature sets for this family exceed 0.86. These feature sets are codon usage plus GC content and tetranucleotide frequencies.

The most prominent members of Orthomyxoviridae are influenza viruses. They cause infectious diseases to mammals and birds spreading out in seasonal epidemics. Orthomyxoviridae are viruses with (-)ssRNA which implies the nucleotide frequencies might deviate from Chargaff’s rules stated in section 2.6.1. The atypical genes identified by both, the labelling inferred from BLAST searches and the one class SVM exhibit nucleotide frequencies in accordance with Chargaff’s rules while a large proportion of genes, mostly from Influenza virus A, show opposing behaviour in the frequencies of A and T. This means, only genes with nucleotide frequency patterns as expected in organisms with dsDNA are classified as atypical for the family of Orthomyxoviridae. The nucleotide frequencies of Orthomyxoviridae genes are depicted in figure 5.16. Atypical genes are only identified within the GC content range of the genes labelled in blue.
Influenzavirus A is the subject of intensive study, and is hence over-represented in the database compared to the other members of this family. For the one class SVM approach, it is possible that this imbalance influences the identification of atypical genes. The BLAST algorithm however is independent of statistical signatures of viral families and yields a similar result: Atypical genes primarily occur among genes not from Influenzavirus A and only in the upper half of the GC content range. Of in total 67 atypical genes identified by the BLAST search, 9 are from Influenzavirus A, while 95% of the Orthomyxoviridae genes in the database are Influenzavirus A genes. One interpretation of these observations is that genes from Influenzavirus A undergo comparatively high selection pressure. While other members of Orthomyxoviridae could take up genes with origins in double stranded organisms, and keep them under comparatively low rates of evolution in the genome, Influenzavirus A genes exhibit high evolutionary rates which have been attributed to positive selection by the human immune system [Nobusawa and Sato, 2006].

Support for our method is provided by the identification of Araguari virus partial glycoprotein (ENA accession ABB55449.1) as an atypical gene. Araguari virus from the Orthomyxoviridae family belongs to the genus of Thogotovirus, and the glycoprotein of Thogoto virus was subject of research of Morse and co-workers [Morse et al., 1992]. It has been found to have resemblances to the baculoviral glycoprotein GP64 indicating a possibility of gene transfer. Interestingly, we did not recover this correspondence in the analysis of similar proteins among unrelated viruses (chapter 4), indicating that the SVM approach is more efficient.
5.3 Results

Flaviviridae For Flaviviridae, AUC scores greater than 0.7 are observed for the feature sets without regression. Flaviviridae have a (+)ssRNA genome and are mainly spread through ticks and mosquitoes. They are subject of intense studies as they are infectious to humans and domestic animals. The majority of Flaviviridae genes in the database stems from Hepatitis C virus. A sub-population of these genes, like Influenzavirus A genes in Orthomyxoviridae, exhibits strong deviations from Chargaff’s rule in their nucleotide composition.

The prediction algorithm as well as the BLAST search based labelling identify atypical genes primarily not from Hepatitis C virus. The first four genes of the ranked result from the SVM which have BLAST hits outside Flaviviridae are from Bovine viral diarrhea virus. To our knowledge gene transfer in these viruses has never been studied. The first of these genes encodes for the non-structural protein NS2-3 (UniProt accession Q65448). For this gene, the BLAST search indicates a weak similarity to Holliday junction ATP-dependent DNA helicase RuvA (J0ZHL7). The closest hit for this protein is found in Bartonella rattimassiliensis. Bartonella species are bacteria infectious to ruminants, and have been studied in cattle vaccinated against bovine viral diarrhea [Maillard et al., 2006]. The fact that both the virus and the bacterium occur in the same environment leaves room for broader investigations which are beyond the scope of this work.

Figure 5.17: Nucleotide frequencies of all genes from Flaviviridae plotted against the GC content. The genes from Hepatitis C virus and Bovine viral diarrhea virus are highlighted.
**Hepadnaviridae** For *Hepadnaviridae*, AUC scores of 0.6 for the estimation based feature sets and scores around 0.8 for the feature sets without regression indicate a reasonable resemblance between the labels inferred from BLAST searches and the prediction result. *Hepadnaviridae* have a dsDNA-RT genome. During their replication process, the genome is integrated into the host genome for transcription. *Hepadnaviridae* infect mammals and birds.

Analysis of the most atypical genes with BLAST hits outside the family contradicts the good AUC scores. The BLAST hits outside *Hepadnaviridae* are due to integration of the virus genome into the host. They stem from a sequence analysis of hepatocellular carcinoma in hepatitis B infected woodchuck [Mitsuyoshi et al., 1991] and from a study on budgerigar genomes being infiltrated by hepadnaviruses millions of years ago [Liu et al., 2012]. The BLAST search in this case provides no information on the validity of the SVM based prediction.

Together with the average AUC scores around 0.5 for all examined families, shown in figure 5.14, the question rises whether the SVM prediction and the BLAST search actually have a reasonable overlap that allows for conclusions. At least for *Hepadnaviridae*, and potentially also for other retroviruses, the labelling inferred from BLAST is not informative and hence allows for no statement about the prediction performance of the SVM.

**Circoviridae** Assuming correctness of the labelling inferred from BLAST, for *Circoviridae* we observe almost perfect predictions with the codon usage feature set. From the first 15 positions in the ranked result, 13 are outliers according to the labels from BLAST. The AUC score for the complete ranking for this feature set is 0.73.

*Circoviridae* are small and have circular ssDNA. They are broadly distributed among vertebrates without causing illnesses for most organisms. The BLAST hits of the top ranked genes refer to a variety of species: *Chicken anemia virus* VP2 (UniProt accession H9AX78) is found to exhibit similarity to the Ribonuclease (M5RQ04) from the bacterial genus *Rhodopirellula*, and to the Potential essential nuclear protein (Q5A077) from the fungus *Candida albicans*. The putative Cap (H9XTK6) is linked to protein GA19527 (accession Q29AM2) from *Drosophila*. Also, the putative Cap (D9IZ83) shows similarity to lptD (F9Q8I5) from the bacterial genus *Haemophilus* and to an uncharacterised protein (I3J755) from the cichlid fish *Nile tilapia*. Furthermore, the Replication-associated protein (Rep) (G1JRP4) and the putative Rep (H9XTK5) link to a Rep-like protein (Q6VZJ4) of *Canarypox virus*, a dsDNA virus.

In the analysis of similar proteins in unrelated viruses, in section 4.3.2, we observed that the circoviral Rep protein exhibits similarity to P41 of *Norovirus* ((+)ssRNA). We do not recover this specific match in the first 15 positions of the ranked result. However, the multiple similarities of the Rep protein to proteins from unrelated viruses open up for hypotheses on the complex history of this protein. It appears to have originated through recombination, combining gene segments from unrelated viruses [Meehan et al., 1997]. Gibbs and co-workers identified Rep-like genes by means of a database search and speculate that these genes originated by multiple interspecies recombinations as they are represented in viral, plasmid, bacterial, and parasitic protozoan genomes [Gibbs et al., 2006].

**Siphoviridae** For some virus families, AUC scores of at most 0.5 indicate poor prediction results. *Siphoviridae* is one example where scores below the performance of a random classification are observed. The reason for these scores is that almost every gene from *Siphoviridae* yields BLAST hits outside the family and thus, to attain a good score, the few genes with no hits outside would need to be at the very end of the ranked result and not somewhere in the...
5.4 Discussion

middle. Our method however is tuned for the identification of minority of atypical genes and not a minority of typical genes. The test with the BLAST search appears to be not applicable in this case.

Siphoviridae are bacteriophages. They infect only bacteria and can carry bacterial genes between them. This mechanism, known as transduction, is one option how horizontal gene transfer among bacteria is accomplished. Therefore, it is not surprising that many genes from Siphoviridae have bacterial counterparts which are detected by the BLAST search. Once more, this shows that the comparison to the BLAST search has drawbacks.

Summing up, the prediction quality on real data is hard to assess. Good or bad AUC scores are neither a clear sign of failure nor of success of the prediction algorithm as they associate two labelling procedures and neither we can validate. Analysis of individual genes of the viral families indicates that the prediction result can be useful to identify potential candidates for HGT. However, the prediction cannot replace a deeper investigation on individual genes. It only provides a hint where to begin with the search for horizontally transferred genes.

It should be pointed out that we cannot expect to find a perfect validation strategy on the true data. Because if this strategy would exist, it could be used in place of the one class SVM and would already solve the problem.

5.4 Discussion

The presented one class SVM model is suitable for the detection of potential examples of horizontal gene transfer. Given statistics about the composition of a viral family, it robustly identifies the genes that are most atypical for the family. Whether the origin of a gene in a family is HGT is a matter of interpretation. The model can provide no guarantee for the identified genes to be horizontally transferred. Its prediction is based solely on statistics using no external information to attain additional evidence for or against the result. The interpretation of the result remains to be derived manually.

Employed Data How usable the presented method is depends also on the data it is to be used for. We did not aim to present an algorithm designed to work nicely on artificially created data sets. Rather, we want to contribute to making sense out of the large amounts of sequence data from viruses available. Unlike artificial data, these data have shortcomings. Aside from issues of correctness and completeness of the gene records, statistical biases are what makes prediction very difficult. While for some viruses there are only one or at most a few genes in the database, others are the object of intensive study and thousands of gene records exist. For a virus family of which one member is highly over-represented, the prediction algorithm will regard the genes of this member as typical for the family provided they have some common statistical features. A solution would be to sample down the genes per species to a common size. This sampling however would imply a vast loss of information which is needed to derive reasonable statistics for the family.

Due to the use of actual data, not only the assessment of performance in the real setting is difficult but also the simulated data sets are affected by shortcomings in the data as they are constructed from these data. The simulated genes added to each family are taken from the set of all other viral genes. As the choice of the simulated genes is uniformly at random with respect to the total number of genes not from the family of interest, it is biased towards well
studied families with a lot of gene records in the data base. Thus, it is possible that the observed performance of families with similar properties as large families is underestimated while the performance found for families with properties very different from the characteristics of large families is overly optimistic. Also, we cannot expect to observe perfect prediction performance on the simulated data. The simulated data is created such that artificial outliers are added to the set of all genes of a viral family. This means, the true unlabelled outliers remain in the data set and should be identified by the prediction algorithm just like the artificial atypical genes.

The comparison to the result of BLAST searches also does not provide a perfect strategy to assess performance due to the biases in the data. Even worse, the BLAST result is affected by the presence or absence of information as well. The effect is different though. BLAST does not construct a statistical model for viral families but summarizes which other genes similar to the one queried exist in the UniProt database. Obviously, if the gene of interest is from a well studied family, very good matches will occur among genes from the same family, occluding matches from other families.

Despite these difficulties, we observe overall reasonable recovery of simulated atypical genes in the data sets and were able to identify genes proposed to be horizontally transferred in the literature. As viruses are highly diverse, with the presented methodology, it is likely that atypical genes in non-viral organisms can be identified as well. A study on atypical genes in archaea or bacteria using the presented method could be the subject of future investigations.

**Modifications and Improvements** The SVM model is build upon a number of decisions which leave room for improvements or at least further investigation. We decided to fix the parameter $\nu$ after analysing the stability of the ranked result to $\nu = 0.2$. Possibly, the algorithm could be extended to find a suitable $\nu$ automatically by evaluation of the silhouette values for different choices of $\nu$. The same holds for the choice of $\gamma$, which we set according to a rule of thumb to $\gamma = (2D)^{-1}$. Conducting a more elaborate selection might improve the prediction result.

Furthermore, we decided to standardize all feature sets noticing that standardization on the sets derived by linear regression effectively shrinks the dimensionality of the feature sets to $D = 2$. Omitting this preprocessing step causes features with high absolute values to be considered more important than others in the prediction algorithm. An alternative is to normalize the data. While standardization transforms the data to have zero mean and unit variance, normalization scales all variables into the range $[0, 1]$ by $x' = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$. Normalization however, is not free of drawbacks. The major drawback of standardization is that it makes the assumption that the data have been generated with a Gaussian law with a mean $\mu$ and a standard deviation $\sigma$ such that the transformation may be computed as $x' = \frac{x - \mu}{\sigma}$. Normalization is free of that assumption but often scales the normal data into a very small interval when there are outliers present. The slope coefficients in the feature sets derived by linear regression would also become unified under this transformation as the minimum value is always 0 and the maximum value always 1 allowing the slopes to be either exactly 1 or $-1$.

The observation that tetranucleotide frequencies yield comparatively good results rises the question if longer oligonucleotide frequencies can even increase the performance. Longer oligonucleotide frequencies could reveal higher order motifs yielding improved results. One issue with them is that the dimensionality increases very quickly making it infeasible to calculate them for large amounts of data in acceptable time. Also, a very large number of samples is needed to cover the space adequately. Tetranucleotide frequency feature sets already have
5.4 Discussion

256 dimensions. Increasing the length by one yields 1024 dimensions. The smaller of the viral families would not provide enough occurrences for all these k-mers to build reasonable statistics. One could also examine the benefit of constructing a feature set incorporating individual dimensions from oligonucleotide frequencies of variable length which are most predictive.

Variations of feature sets have been considered intensely, but we did not pose the question, how suitable a one class SVM model is in comparison to other methods, such as clustering or HMMs.

Also, we did not explore the effect of changing the taxonomy level at which the viruses are considered in this study. When families exhibit high levels of diversity, the use of family level might be inappropriate, and the use of one statistical model for a whole family becomes a questionable approximation. Alternatively to predicting atypical genes for virus families, one could predict on the level of species. Members of *Myoviridae* for instance, can infect Bacteria and Archaea. As there is evolutionary pressure towards adaptation to the host [Bahir et al., 2009], codon usage pattern of the species infecting Archaea assumably are different from the pattern of those infecting Bacteria. On the other hand, viral species which are not the subject of intense study can be taken into account using family level whereas on the level of species there would not be enough data available to derive any valuable information for these species.

Related Results  Alternatively to regarding genes as a unit of information, one could use a sliding window approach on complete viral genomes. Together with switching to species level, this modification would allow to compare our results to other published work. Tsirigos and Rigoutsos, for instance, devised a sliding window approach to identify atypical sequences in bacterial, archaeal, and viral genomes [Tsirigos and Rigoutsos, 2005b], stressing the great sensitivity of their approach. A comparison of performance would be interesting as in this approach, a one class SVM model has been employed as well. Details in the use of this model however differ. While we fix parameters beforehand and report the ranking induced by the distances of each gene to the decision hyperplane, Tsirigos and Rigoutsos estimate the parameter $\nu$ such that, in an artificial setting, the ratio of recovered artificially inserted genes is maximized. As this kind of selection is infeasible in a real setting, regions are reported as a candidate for gene transfer only if they were marked as atypical at a number of tested values of $\nu$. Their result is binary reporting either evidence for HGT or against for every part of a genome. One may pose question if the performance, assessed primarily on the artificial data, is comparable to the performance on real data, as the parameter selection processes used in both settings are very different and the SVM result heavily depends on this parameter. Apart from differences in the implementation of the model, the major distinction of our approach is that we use no sliding window but consider continuous coding regions, because we wanted to analyse the transfer of complete functional unit such as proteins.

A comparison to results of a study of evolution of the family of *Poxviridae* [Odom et al., 2009] could prove valuable for the assessment of performance of our approach. Odom and co-workers investigate the evolutionary origins of proteins encoded by the *Poxviridae* family using a method of characterizing and visualizing the similarity between these proteins and taxonomic subsets of proteins. They use taxonomic group plots allowing to visually compare two sets of BLAST scores obtained from querying mutually exclusive taxonomic databases for a set of proteins.

A future direction could be to adapt the devised algorithm for the identification of HGT to work for non-viral organisms as well. This would allow for more extensive comparison to other results for well studied cellular organisms.
6 Conclusion

Two systematic analyses of HGT in viruses have been presented in this work. With both these analyses together, we provide a broad overview of related proteins in distant viral genomes. We systematically show what unrelated viruses have in common and provide a method to identify what is atypical in a virus family. These results might contribute to the broad picture of virus evolution and the role of viruses in the evolution of life.

In the first part, we analysed the similarities of proteins from unrelated viruses. We find proteins of all virus types to have counterparts in viruses from at least one other virus type. The proteins we most commonly identify as similar between different virus types belong to the replication machinery and some are ubiquitously present also in cellular organisms. For most identified similar proteins, evidence for their homology is strengthened also by other literature. Contrary to our initial hypothesis, no clear indication for HGT between unrelated viruses can be provided. The analysis of GC contents however provides evidence for some proteins to be transferred from the host into the viral genome.

In the second part of this work, we aimed to identify viral proteins that have alien origin. By evaluation of statistical signatures of all known viral genes, employing a one class SVM algorithm, we could rank proteins according to how typical they are in the virus family. The evaluation of the presented SVM based prediction algorithm included a comparison to the result of a BLAST search. We analysed in detail proteins from selected virus families that are both predicted to be very atypical by the SVM algorithm and found to exhibit similarities to proteins not from their own viral family by the BLAST search. For some of these proteins, support of our prediction can be found in the literature. Others reveal shortcomings of the validation approach: For retroviruses, for instance, a large fraction of the BLAST hits stems from infected organisms. These hits show examples of gene transfer from the virus to its host, but we aim to study gene transfer into viruses. Because these hits are numerous and yield high similarity scores in the BLAST search, they possibly conceal information more relevant to this analysis.

In the future, we hope to extend the study of homologous proteins in unrelated viruses, incorporating structural features of proteins. The conservation of protein structures might help to discover relations otherwise hidden through the rapid evolution of viral proteins, which makes relatedness undetectable on sequence level.

We plan to make the method to detect atypical proteins in a virus family publicly available as a web tool. This tool would enable the possibility of automated classification of proteins newly added to UniProt, indicating whether they are typical for the assigned virus family or not. Furthermore, the tool could be generalized to be usable with non-viral data.
List of Figures

2.1 The universal genetic code ............................................. 5
4.1 Virus types with matching proteins ................................... 19
4.2 Active site of Mycobacterium tuberculosis dUTPase ................ 21
4.3 Alignment of dUTPase proteins ......................................... 22
4.4 Alignment of RNA helicase proteins ................................... 23
4.5 Structure of Bacteriophage HP1 integrase ............................. 25
4.6 Alignment of phage integrase proteins ................................. 25
4.7 Alignment of the C-terminal domain of helicase superfamily 2 proteins .......................................................... 27
4.8 Alignment of haemagglutinin esterase proteins ...................... 28
4.9 Receptor binding site of HEF ............................................. 29
4.10 Alignment of RNA dependent RNA polymerase proteins ........... 30
4.11 Structure of RNA dependent RNA polymerase ..................... 31
4.12 Alignment of serine/threonine-protein kinase ....................... 32
4.13 Alignment of OrfB IS605 proteins ....................................... 33
4.14 Alignment of Parvo NS1 proteins ....................................... 35
5.1 Decision boundary of a one class SVM using different values of $\nu$ ........................................................................... 46
5.2 Decision boundary of a one class SVM using different values of $\gamma$ .................................................................................. 47
5.3 Linear dependence between GC content and nucleotide frequencies ................................................................................. 48
5.4 Linear regression of position based nucleotide frequencies ........ 49
5.5 Behaviour of MCC scores for different class proportions .......... 53
5.6 Ranking stability under varying $\nu$ for different values of $\gamma$ ....................................................................................... 57
5.7 Distribution of the number of genes per family and the success rates in each bin ................................................................. 59
5.8 Success rates of the one class SVM for different values of $\nu$ ........ 60
5.9 Success rates of the one class SVM for different feature sets .......... 61
5.10 Distribution of GC content ranges per family and the success rates in each bin ................................................................. 63
5.11 Prediction success per virus type on different feature sets .......... 64
5.12 Boxplot of AUC scores on different feature sets ..................... 65
5.13 Boxplot of silhouette values on different feature sets ................ 66
5.14 Boxplot of AUC scores on selected feature sets feature sets .......... 68
5.15 ROC plot for the prediction of atypical genes in Orthomyxoviridae on real data ................................................................. 69
5.16 Nucleotide frequencies of Orthomyxoviridae genes plotted against GC content ................................................................. 70
5.17 Nucleotide frequencies of Flaviviridae genes plotted against GC content ................................................................. 71
List of Tables

4.1 Structural families in unrelated viruses ........................................ 20
4.2 Proteins with largest deviations in GC3s content from the corresponding genome 36
4.3 Proteins with GC3s content similar to that of the matching protein ........... 37
4.4 Proteins with GC3s content similar to that of the host .......................... 38
List of Acronyms

HGT  Horizontal gene transfer
LGT  Lateral gene transfer
SVM  Support vector machine
DNA  Deoxyribonucleic acid
RNA  Ribonucleic acid
A    Adenine
C    Cytosine
T    Thymine
G    Guanine
U    Uracil
ICTV International Committee on Taxonomy of Viruses
UniProtKB UniProt Knowledgebase
UniProt Universal Protein Resource
EMBL-EBI European Bioinformatics Institute
SIB   Swiss Institute of Bioinformatics
PIR   Protein Information Resource
INSDC International Nucleotide Sequence Database Collaboration
CDS   Coding sequences
BLAST Basic Local Alignment Search Tool
HMM   Hidden Markov model
PDB   Protein Data Bank
ENA   European Nucleotide Archive
MCC   Matthew’s correlation coefficient
tp    True positives
fp    False positives
tn    True negatives
fn    False negatives
SD    Standard deviation
dsDNA Double-stranded DNA
ssDNA Single-stranded sense DNA
dsRNA Double-stranded RNA
(+)-ssRNA Single-stranded sense RNA
(-)-ssRNA Single-stranded antisense RNA
ssRNA-RT Reversely transcribed single-stranded sense RNA
dsDNA-RT Reversely transcribed double-stranded DNA
AUC   Area under curve
ROC   Receiver operator characteristic
dUTP  Deoxyuridine triphosphate
dUDP  Deoxyuridine diphosphate
dUMP  Deoxyuridine monophosphate
**List of Acronyms**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>dTMP</td>
<td>Deoxythymine monophosphate</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td>GC3s</td>
<td>Third position GC content</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>NTP</td>
<td>Nucleoside triphosphate</td>
</tr>
<tr>
<td>attP</td>
<td>Phage attachment site</td>
</tr>
<tr>
<td>attB</td>
<td>Bacterial attachment site</td>
</tr>
<tr>
<td>Rep</td>
<td>Replication-associated protein</td>
</tr>
<tr>
<td>HEF</td>
<td>Haemagglutinin-esterase fusion glycoprotein</td>
</tr>
</tbody>
</table>
Bibliography


Bibliography


Bibliography


Bibliography


